DOCUMENT RESUME

ED 041 856 24 SP 004 122

AUTHOR	Abramson, Theodore
TITLE	Development of Improved Techniques for Establishing the Reliability of Observation Ratings. Final Report.
INSTITUTION	City Univ. of New York, N.Y.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	BR-9-B-070
PUB DATE	Jan 70
GRANT	OEG-2-9-400070-1039(010)
NOTE	83p.
EDRS PRICE	EDRS Price MF-$0.50 HC-$4.25
DESCRIPTORS	Analysis of Variance, *Classroom Observation Techniques, Measurement Techniques, *Observation, Reliability, Statistical Analysis, *Student Behavior, *Teacher Behavior

ABSTRACT

This investigation sought to develop and apply analysis of variance techniques to the estimation of the reliability of observation schedules. It placed special emphasis on the different possible designs and the various administrative situations in which they might be applied. The application of the general model to a specific instance was then carried out. The study was conducted with ten recorders who observed five teachers through a one-way mirror and rated them on an observational schedule. The SUTEC (School University Teacher Education Center) Observational Schedule was used. (A copy is attached to the document.) Items observed included teacher mobility, the involvement of the children, materials present, materials in use, irrelevant behavior by the children, directed behavior, and spontaneous behavior. Analysis of variance was applied to the data obtained. The major conclusions were that different variance component models could be applied in different situations to estimate the reliability of either the entire observation schedule or parts of it, and that the items comprising the SUTEC schedule did differentiate fairly well between teachers. (Author/MBM)

BR 9-B-070
PA 24
SP

ED0 41856

FINAL REPORT
Project No. 9-b-070
Grant No. OEG-2-9-400070-1039(010)

U.S. DEPARTMENT OF HEALTH, EDUCATION
& WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

DEVELOPMENT OF IMPROVED TECHNIQUES FOR ESTABLISHING THE RELIABILITY OF OBSERVATION RATINGS

Theodore Abramson
School of Education
Fordham University
Lincoln Center Campus
New York, N. Y. 10023

January 1970

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

SP004122

TABLE OF CONTENTS

PAGE

SUMMARY vii

CHAPTER

I. INTRODUCTION. 1

Statement of the Problem. 2

Definition of Terms 3

Significance of the Problem 4

Limitations of the Study. 5

Review of Related Literature. 5

Traditional Methods of Calculating the Reliability of Observational Data . . . 6

Summary of Literature on Traditional Methods of Calculating Reliability of Observational Data. 11

Recent Methods of Calculating the Reliability of Observational Data . . . 11

Summary of Literature on Recent Methods of Calculating Reliability of Observational Data. 23

Summary of Related Literature 24

II. THE SUBJECTS, MATERIALS, AND PROCEDURES . . 25

The Subjects. 25

The Materials 26

The Procedures. 27

The Statistical Procedures. 27

CHAPTER PAGE

The Variables and Designs 31

III. ANALYSIS OF THE RESULTS OF THE INVESTIGATION 55

Reliability of Individual Items 55

Reliability of the Entire Schedule. . . . 59

IV. CONCLUSIONS AND RECOMMENDATIONS 65

Conclusions 65

Recommendations 66

REFERENCES. 68

APPENDIX. 70

LIST OF TABLES

TABLE PAGE

1. Medley and Mitzel Reliability ANOVA 15
2. Medley and Mitzel Expanded Reliability ANOVA 17
3. Medley and Mitzel Estimation of Variance Components. 19
4. Schematic Representation of a Three Factor Partially Nested Design 33
5. Schematic Representation of Two Factor Repeated Measures Design. 34
6. Two way ANOVA for Repeated Measures 37
7. Estimation of Variance Components for a Two Factor Repeated Measures Design 38
8. Representation of Three Factor Repeated Measures Design with n Subscores. 39

TABLE PAGE

9. Three Factor ANOVA with Repeated Measures on One Factor. 41

10. Estimation of Variance Components for a Three Factor Design with Repeated Measures on One Factor 42

11. Schematic Representation of Three Factor Design with Two Repeated Measures and n Subscores. 43

12. Three Factor ANOVA with Repeated Measures on Two Factors 44

13. Expected Mean Squares of Three Factor Design with Repeated Measures on Two Factors. . . 45

14. Estimation of Variance Components for a Three Factor Design with Repeated Measures on Two Factors 46

15. Four Factor Design with Two Repeated Measures 48

16. Four Random Factors Design with Two Repeated Measures 50

17. Degrees of Freedom for a Four Factor Design with Two Repeated Measures 52

18. Variance Components for a Four Factor Design with Two Repeated Measures 53

19. Analysis of Variance of the SUTEC Observation Team Data. 56

TABLE PAGE

20. Estimation of Variance Components of the SUTEC Observation Team Data. 56

21. ANOVA for the Four Observers Present During Both Observations. 57

22. Estimation of Variance Components for the Four Observers Present During Both Visits. 57

23. Variances and Correlations for the Entire Observation Team and the Four Observers Present During Both Observations 58

24. Sources of Variation, Degrees of Freedom, and Expected Mean Squares for an ANOVA Design with Factors B Nested under Factor A 60

25. ANOVA Design with Factor B Nested under Factor A, All Factors Random, and n = 1. . 61

26. Analysis of Variance of an Observation Schedule Containing Seven Items and Using Three Observer Teams and Three Teachers. . 62

27. Estimation of Variance Components for an Observation Schedule Containing Seven Items and Using Three Observer Teams and Three Teachers 63

Acknowledgments

Most of this report resulted from the author's dissertation in the School of Education at Fordham University. The work was conducted under the mentorship of Dr. Frances J. Crowley whose help and guidance are gratefully acknowledged.

Summary

Innovations in teacher training are often dependent on observational data. The problem of measuring reliability of observations collected by a team is due to (1) the difficulties of maintaining an observer team intact over an extended period of time and (2) observing each teacher a number of times, or more than once. These two conditions are normally required if one is to apply the Analysis of Variance (ANOVA) model proposed by Medley and Mitzel in 1963.

This study presented a number of different ANOVA models and the administrative conditions under which they were to be applied such that the partitioning of the sources of variation and the calculation of reliability coefficients could be carried out. Specifically, a model was designed for the observer team situation in which the team visited a number of different teachers only once and where the team did not necessarily contain the same members for all visits. The paradigm was developed for situations in which there were n observations per item per observer and also for the situations when there was only one observation per item per observer.

The model was applied to data collected by teams of observers from the use of an observation schedule of teacher and pupil classroom situations and behaviors. The schedule items were teacher mobility, involvement of children, materials present, materials in use, directed behavior, spontaneous behavior, and irrelevant acts.

The reliabilities of the mobility, involvement of children, and irrelevant acts were .72, .67, and .69, respectively. The overall reliability coefficient of .37 and the variance components of .38, .18, and .07 for the items, interaction, and error terms respectively indicated that the teacher and item factors accounted for 75% of the total variance.

Future research which would field test and compare different administrative situations and their respective reliability coefficients calculated from the appropriate designs was recommended.

CHAPTER I

Introduction

Teaching is often considered an applied science or art. This concept would lead one to expect that a good deal of research on teaching has been done by observing the classroom where the underlying educational principles are actually applied. Medley and Mitzel, in referring to the paucity of such research, stated:

> Certainly there is no more obvious approach to research on teaching than direct observation of the behavior of teachers while they teach and pupils while they learn. Yet it is a rare study indeed that includes any formal observation at all (Medley & Mitzel, 1963, p. 247).

In recent years a number of different classroom observation schedules which permit the classification of teacher and pupil behaviors into a variety of category schemes have been formulated. Although not explicitly stated in most instances by their originators, the underlying rationale for many observation schedules was that given by Soar who stated ". . . it is possible to identify and measure a common core of teacher-pupil classroom behaviors which are basic to most, if not all (important) aspects of pupil intellectual, personal, and social growth"(Soar, 1966, p. 2). Medley (1967) indicated that the theoretical formulation behind the construction of the Observation Schedule and Record (OScAR) consisted of the relationship between three levels of teacher behavior and effectiveness. The levels consisted of the variables related to classroom climate, the conducting of learning experiences, and the maintaining of pupil involvement.

The advent of these schedules has actually made possible the increased application of research technology to a wide variety of educational problems such as school program evaluation and teacher preparation program evaluation. The schedules have also facilitated the development of theories of teaching. Certainly, data drawn directly from actual classroom behavior provide a more adequate sample of the teaching-learning situation from which inferences can be drawn on the worth of a program, than do such ad hoc factors as pupil achievement or attitude toward the "new" program.

Observational data, unfortunately, besides being expensive to obtain, is no more precise an index of actual behavior than the team's ability to observe and classify accurately that which transpires in the classroom. Therefore, what is required is not merely a well trained team, but a team whose members see and report the same things with accuracy and consistency so that in effect the data reported by different members of the team are comparable. To insure maximum comparability and minimum variation of data collected by the members of the observation team a schedule is usually devised. The schedule, by listing the cues to be responded to, helps to minimize the observer error. Thus the usefulness of observational data is to a great extent dependent on what has been called inter-observer agreement or reliability.

In the past the reliability of observation schedules has usually been examined through correlation analyses which yielded a measure of inter-observer agreement between only two observers. During the five years from 1958 to 1963 an analysis of variance (ANOVA) technique (Medley & Mitzel, 1958a, 1963) using a factorial design was developed which permitted the variance to be partitioned into its component parts and the calculation of an overall reliability coefficient. The application of this model required that the same observers visit the same teachers a number of times. The logistical problems involved in applying this ANOVA model have made it administratively unfeasible and to date very little use has been made of this method of calculating reliabilities. What was required in order to make the ANOVA technique more applicable? What assumptions and restrictions would have to be imposed if the technique were to be statistically valid? These were some of the basic questions which this study addressed itself to and sought to answer.

Statement of the Problem

The purpose of the study was to investigate the conditions under which an ANOVA model or models could be applied to the calculation of an overall reliability coefficient and the partitioning of the sources of variation into its component parts of an observational schedule without requiring the same administratively unfeasible conditions as the original model. The general models were then applied to the data obtained on the School University Teacher Education Center (SUTEC) Observation Schedule. The purpose of this application was to make explicit the steps that were required in order to apply the more general model to a specific instance.

To this end, the study investigated and endeavored to answer the following questions:

1. Which variables and interactions between variables were to be expected when dealing with observation schedules that were, and may continue to be, used to study classroom behavior?

2. What was the difference between a "random" and a "fixed" factor as far as the model was concerned? Which of the variables identified in question 1 were "fixed" and which were "random?"

3. Was it possible to consider the original factorial model (Medley & Mitzel, 1958a, 1963) as a repeated measures design? If so, what difference would this make in the general analysis?

4. Under what conditions could designs different from the original design be developed and applied to make them more administratively feasible?

5. What was the relationship between analysis of variance and the reliability of an observational schedule defined in terms of the variables and sources of variation inherent in an observation schedule?

6. What assumptions had to be made to make possible the application of the general model or models to the specific observation schedule data available--namely, the SUTEC data?

7. In general, how can the reliability coefficient and the variance components be used to estimate the percentages of the variance attributable to the factors involved?

Definition of Terms

A number of terms employed during this investigation required definition. However, the more technical terms which pertained to the ANOVA models such as "random," "fixed," "finite," "crossed," and "nested" factors are discussed and defined in the section of Chapter II on Statistical Procedures. The more general and less statistical terms that needed to be defined were: observation schedule and observation team.

Observation Schedule. For the purposes of this investigation the term observation schedule was defined as a series of selected items that categorized and/or described those classroom behaviors of teachers and students and/or settings to which trained observers were directed to attend. The items were typically formalized into a category scheme and prepared in a form (list, grid, etc.) which permitted rapid recording of observations.

Observation Team. For the purposes of this investigation observation team was defined as the group of people who were trained to collect classroom data through the use of an observation schedule.

Significance of the Problem

It has been indicated that an ANOVA model to calculate reliabilities of observation schedule data has been formulated. This theoretical approach, specifically geared to the variables present in a classroom observation situation, was first proposed, and subsequently further developed by Medley and Mitzel (1958a, 1963). However, to date very little use of this ANOVA model has been reported in the literature.

It is believed that the general lack of application of this model in the past to studies involving observation teams was in large measure due to the practical difficulties involved in its application. Maintaining an observation team intact over an extended period of time and being permitted to visit the same classrooms and teachers a number of times is difficult and rather expensive. Both of these conditions, however, using only the same observers and the same teachers, are necessary if one is to apply the previosly discussed ANOVA model. Practical research administration problems usually force the researcher to train his team by having all the members of the team visit a classroom together after they have become somewhat familiar with the observation schedule that they will use. Subsequent to the first visit a group discussion is then typically followed by a visit to another teacher. This procedure is followed until there is a fair amount of agreement between observers at which point the members of the team are sent out individually to observe the teachers who are the Ss of the study. The presently available ANOVA techniques do not apply to analyses of observer team data under these frequently prevailing conditions.

This study therefore was devoted to the development and application of an ANOVA model or models applicable to conditions when the observer team did not necessarily have the same observers and when the observation of a given teacher did not necessarily occur many times. The assumptions and procedures necessary in order to apply the general model to a specific case were also made explicit. The development of the model makes possible the broader and more precise use of observation team data in a variety of educational problem situations and thus makes feasible more accurate appraisals and evaluations than are currently possible. At the same time, the availability of another method of measuring the reliability of observation data may aid in directing a greater research effort to the place where the teaching-learning process is carried on--the classroom.

Limitations of the Study

This study was limited with respect to the following factors:

1. Applicability. The inherent complexity of the subject puts the reliability calculation beyond the present training of many research workers, although the administrative requirements have been simplified. Therefore, the anticipated greater applicability of observation data, in general, and the ANOVA model, in particular, may not occur.

2. Validity of the Schedule. The models developed in this study dealt only with the problem of the reliability of observation schedules. At the same time the sources of variation, expected mean squares, and the percent of variance attributable to various sources were calculated. However, the foregoing in no way answers questions pertaining to the appropriateness, usefulness, validity, etc. of the items comprising the schedule. Therefore, the validity of the SUTEC Observation Schedule is still highly questionable and although the calculated reliabilities may be numerically acceptable they may be meaningless. This limitation is clearly a function of the construct validity of specific schedules and is directly related to the underlying rationale of each schedule and its originator's philosophy or theoretical framework and is therefore beyond the scope of the present investigation.

Review of Related Research

During the last 20 years a number of observation schedules which purport to measure classroom behaviors and/or settings have been formulated and used. Many of the newer instruments such as the Observation Schedule and Record (OScAR) developed by Medley and Mitzel (1958b) may be considered as refinements or amalgamations of parts of previously proposed schedules. OScAR was actually based on the category schedules of Withal and Cornell (Medley & Mitzel, 1958b, 1963) and was supposed to measure emotional climate, verbal emphasis, and social structure. A thorough review of the many category schemes and their uses for the period up to 1963 is available (Medley & Mitzel, 1963) and need not be repeated.

In developing or using an observational schedule a problem that must be faced is that of the reliability of the data. The problem of the reliability of observation data has usually been treated in terms of the per cent of observer agreement or in terms of an interclass correlation between two sets of observations. In the latter case a Pearson r or a rank order coefficient has usually been used. Not only was this so prior to 1963 (Medley & Mitzel, 1963), but a review of the more current literature seemed to indicate that this was still essentially true.

For the purposes of this review the studies dealing with the reliability of observation data that were investigated were grouped under two main headings: traditional reliability calculations, and other methods of calculating reliability. The reports that used the traditional methods will be considered first and will be followed by the studies which used methods other than per cent of observer agreement and/or the Pearson Product Moment coefficient of correlation or its equivalent parametric or non parametric counterparts. Studies which did not involve classroom situations were included in the review since the techniques of measuring reliability were the central issue rather than the content or discipline of the application. This sectioning was done to provide a frame of reference and when a study fitted into both categories this categorization was not strictly adhered to.

Traditional Methods of Calculating the Reliability of Observational Data

The review cited earlier (Medley & Mitzel, 1963) was not only replete with the various observation instruments up to 1963 but also contained fairly complete information on the reliability calculations of observation data up to that time. Because of the great familiarity of many research workers with the traditional concepts of reliability and the lack of any additional contribution which might ensue from a second review of the period up to 1963, this section was devoted only to studies reported after 1963.

Ojemann and Snider (1964) reported on the development and scoring of an observation form that was to evaluate part of the Preventive Psychiatry Research Program of the University of Iowa. The aim of the program that was being evaluated dealt with the construction of curricular materials that would help children acquire an understanding and appreciation of the dynamic nature of human behavior. That is, the materials were to help the children develop a causal approach to their social environment. To fulfill this aim a teaching program in "behavioral science" was constructed.

The observers who were to do the observing were trained during the spring preceding the form's actual use. The training consisted of a group discussion of the items comprising the observation schedule, preliminary observation of three children by each observer, and subsequent group discussion and clarification of the items.

Following this training a study of the reliability of the instrument was carried out. Two observers, I and II, carried out simultaneous observation on 32 fourth grade Ss. Similar observations were conducted by observers II and III on a class of 28 fifth grade Ss. The correlation between the behavior scores of the two observers were .69 and .67 for the fourth and fifth grade Ss respectively.

Rusch, Denny, and Ives (1964) reported on the development of a test of creativity in the dramatic arts. Part I of the test was objective and purported to measure fluency and redefinition. The S listened to a tape-recorded story and was shown a piece of cloth. The S then listed the number of ways the material might be used in putting on a play of the story. Fluency was determined by the number of responses given and redefinition was scored as the number of unusual uses suggested for the object. A similar procedure was used for a piece of driftwood and some mood music.

Part II of the test consisted of the S's writing a short description of how he would produce a scene from the story he had heard. Part II was rated on a five point scale for originality and sensitivity by three independent evaluators. Two groups from two sixth grade classes were matched on IQ, reading achievement, and sex. The 47 eleven and twelve year olds were given alternate forms of the test in the fall and spring of 1959. The reliability of Part I of the test ranged from .38 to .67. The reliability of Part II, which required value judgments on the part of the raters, was given in terms of per cent agreement. Per cent of agreement was given in terms of raters I and II, I and III, II and III, and ranged from 8.7% to 69.0% for sensitivity and 2.2% to 66.7% for originality.

Courson (1965) was interested in whether inference as a technique for gathering data was an acceptable research tool. He indicated that this problem was essentially a question of the reliability of the data. To test his hypotheses about interscorer reliability and stability of data based on inferences he studied the reliability of inferences of trained observers on samples of simulated behaviors. The Ss were 64 high school seniors each of whom wrote an assigned projective essay on the topic of "A Teenager's Advice to the World" and an essay in response to the simultaneous presentation of Cards 1, 4, and 20, of the Thematic Apperception Test. The three raters were trained in the use of the nine point Perceptual Factors Rating Scale which contained the three items:

1. How does this person see himself?
2. To what extent is this person identified with others?
3. To what extent is this person open to his experience?

Each observer's ratings were correlated by calculating a Pearson coefficient. A month after the ratings were completed each observer rescored a sample of 10 of the essays and a correlation between initial and final ratings was calculated. The interscorer r's ranged from .38 to .55 and the stability coefficients ranged from .72 to .84.

Maas (1965) investigated the reliability of adjective rating scales. A scaled expectation rating scale was constructed as follows: First, a committee of interviewers who were familiar with the job to be performed established the traits that were to be performed. Second, examples of on the job behaviors which illustrated high, average, and low degrees of the trait were written. Third, the traits were reallocated back into traits and levels by independent judges. Fourth, only those examples with complete agreement as to trait and level were retained. Finally, the remaining examples were arranged on a continuous vertical scale with each example at its proper scaled level for the trait. An interview guide with the weights for each trait at the back of the trait pages was prepared for use with the scaled ratings.

Maas then compared the reliability of the traditional adjective rating scale with the scaled expectation rating method and found significant differences in favor of the latter. During the first year of study 360 Cornell University undergraduates were interviewed twice in the traditional manner. The questions on the second interview differed slightly from those on the first interview, and therefore the inter-interviewer reliability actually consisted of inter-interviewer and S reliability. The

Pearson correlations were .35 for the trait scores, .34 for the overall rating, and .34 for the grand total score. The following year, using the same procedures, 500 candidates were interviewed using the scaled expectation rating technique. The correlations this time were .58, .47, .55. Subsequently, the scaled rating technique was used by three interviewers, interviewing 188 and 172 candidates for female and male dorm counselors, respectively. When interview guide and candidate reliability were held constant, inter-interviewer reliability was found to be .69 for the trait scores, .65 for the overall rating, and .72 for the grand total.

Bobbitt, Gordon, and Jensen (1966) studied the continued inter-observer agreement of pairs of observers for a two and one-half year period. The data were collected on three groups of four mother-infant pairs of pigtail monkeys for five random 10 minute samples of behavior per week for 26 weeks. A pair of observers simultaneously scored the behavior of mother and infant for one of the 10 minute periods per week. Prior to this study the observers were required to attain a .75 agreement percentage criterion where agreement percentage is equal to the ratio of the responses agreed on to the total number of responses. Following each observation an agreement percentage was calculated and a discussion of the observation was held. The dimensions measured were position, posture, locomotion, visual, oral, and manipulation. For all groups the total agreement percentage was .79 with all dimensions ranging from .75 to .89 except for the visual dimension which was .55.

Zunich (1966) studied the relationship of child behavior and parental attitudes of 18 boys and 18 girls whose ages in years and months ranged from 2-9 to 5-0. Direct observation of the children, utilizing a time sampling technique of five minute duration and predetermined categories, was conducted through a one-way mirror. The categories observed were: asking permission, contact, cooperation, criticism, directing, indications of anxiety, interference, non-cooperation, playing interactively, praise, remaining out of contact, restricting, seeking attention, seeking contact, seeking help, seeking information, seeking praise, and suggests.

Reliability of the observations was calculated in terms of the percentage of agreement between two observers who recorded the behaviors of children who were not included in the study. Behavior was recorded simultaneously and independently by two observers during 30 five-minute periods with an observation being made every five seconds.

The number of agreements divided by 60, the total number of observations for a five-minute period, was equal to the percentage of agreement. For the 150 minutes of observations, the 30 five-minute periods, the reliability ranged from .83 to .97.

Bloom and Wilensky (1967) constructed an observation scale to measure the behavior of teachers. The scale was based on a Skinnerian framework and contained the following four categories: information giving, response elicitation, feedback, and teacher control.

The Ss of the study were 72 underprivileged nursery children. Each observation lasted five minutes and was prorated if the activity ceased during the observation. For each of the observational categories, the inter-rater reliabilities, based on 26 five-minute observational periods exceeded .90.

In the development of the Behavior Survey Instrument, an observation sheet geared to Head Start and other special programs in early childhood education, Katz, Peters, and Stein (1968) used the agreement percentage as their measure of reliability. An overall agreement percentage of 84.6% was attained by simultaneous independent observation of the same Ss and was based on the seven categories which were: task orientation, satisfaction, motivation, cognitive, motility, interpersonal behavior, and situation. The range for the categories was .64 to .98.

Brown, Mendenhall, and Beaver (1968) developed the Teacher Practices Observation Record (TPOR) which attempted to measure the agreement of teachers' observed classroom behavior with educational practices that were advocated by John Dewey. The TPOR had seven categories and contained a total of 62 items. The categories were: nature of the situation, nature of the problem, development of ideas, use of subject matter, evaluation, differentiation, and motivation control. Five filmed lessons were observed in 1964 by 130, 124, 119, 119, and 67 observers who received only a 10 minute explanation of the instrument. The observers were drawn from two large midwestern universities and east coast and west coast teacher training institutes. The observer judges were occupationally, college supervisors of student teaching, education professors, and academic professors. No significant differences were found between any of the groups on films 1, 2, 4, and 5. There was a significant difference on film 3 between supervisors of student teaching and both education and academic professors.

In 1965 films two and four were observed once again by 69 and 72 of the judges. Pearson coefficients for the observers' total scores within a given viewing and for the repeat viewings were calculated and ranged from .86 to .93 and .27 to .65, respectively. Correlations for each 10 minute segment of each 30 minute film were also given and ranged from .52 to .71.

Summary of Literature on Traditional Methods of Calculating Reliability of Observational Data

In summarizing the research reported in this section, one must be mindful of the fact that each of the studies used a different research design. This was as it should be, since each investigation was essentially considering a different problem. There were differences in the instruments employed, the number of subjects who participated and the hypotheses being tested.

With the exception of one study (Brown et al., 1968), which will be considered again in the next section of this report, all of the studies reviewed were concerned with the reliability of their observational data as a secondary problem. Because other problems were of primary importance, reliability considerations were often treated in a superficial fashion. These studies all had in common their traditional method of calculating reliabilities, i.e. percentage of agreement or Pearson r, or their equivalents. The question of whether or not these methods or reliability calculation were the best ones available, or should even have been employed were for the most part ignored or at best cursorily treated.

Recent Methods of Calculating the Reliability of Observational Data

Scott (1955) developed an index of interscorer agreement, Pi, for nominal scale coding. That is, Pi was to measure interscorer agreement when the coding dimensions were not ordered along equal intervals or along a dimension of "more or less" of some attribute. The index, Pi, was to be used in survey and observational research where the typical procedure had usually called for one coder to analyze and code interview data. The data were then categorized by a second rater, and then a comparison between the two analyses was made. These analyses were followed by a conference between the two raters to enable them to arrive at their "best" judgment.

The value of Pi was equal to the ratio of the difference between the percentage of actual agreement and the

the agreement expected on the basis of chance to the difference between maximum chance agreement and agreement expected on the basis of chance.

Symbolically,

$$\underline{Pi} = (P_o - P_e)/(1 - P_e)$$

where P_o was the percentage of agreement between the two independent analysts, and P_e was the per cent of agreement to be expected on the basis of chance.

The expected per cent agreement for the dimension was equal to the sum of the squared proportion over all categories.

Symbolically,

$$P_e = \underset{i=1}{\overset{k}{\mathrm{SUM}}}(p_i^2)$$

where k was the total number of categories and p_i was the proportion of the entire sample falling into the i^{th} category.

As an illustration of the method Scott (1955) calculated the value of *Pi* for the question "what sorts of problems are your friends and neighbors most concerned about these days?"

Nature of Problem	Per Cent of All Responses
Economic problems	60%
International problems	5%
Political problems	10%
Local problems	20%
Personal problems	3%
Not ascertained	2%

Therefore, $P_e = (.60)^2 + (.05)^2 + (.10)^2 + (.20)^2 + (.03)^2 + (.02)^2 = .41$. On the basis of an assumed 80% agreement between observers, the index of inter-coder agreement was:

$$\underline{Pi} = (.80 - .41)/(1 - .41) = .67$$

In his discussion of observer reliability for his verbal category system, Flanders (1960) estimated interobserver agreement through an adaption of the Scott (1955) coefficient, *Pi*. Rather than actually using the formula to calculate P_e and *Pi*, he (Flanders, 1960) developed approximations to P_e which were based on graphic estimates.

The graphic estimate of P_e was then followed by a graphic estimate of Pi.

Two observers were trained to use the Flanders system which contained seven teacher, two pupil, and one general category as follows: Teacher accepts feeling, teacher praises/encourages, teacher accepts/uses ideas, teacher asks questions, teacher lectures, teacher gives directions, teacher criticizes/justifies, student responds, student initiates, and silence/confusion. The proportion of tallies of the observers in each category was found and was used to calculate P_e. The value of P_e was also estimated graphically for both observers. The values of Pi using the calculated and estimated value of P_e were .855, .853, and .854, respectively. A critical ratio comparing these values was not carried out, although such a critical ratio calculation was possible (Scott, 1955), because it was obviously unnecessary.

Furst and Amidon (1967) used Flanders' interaction analysis to investigate differences in interaction patterns between elementary school teachers of different subjects and of grades one through six. One hundred sixty classroom observations, one-third in "ghetto," one-third in suburban, and one-third in urban "middle" socioeconomic level schools, were carried out. There were a minimum of 25 observations at each grade level with at least five observations in the areas of arithmetic, social studies, and reading at each grade level.

The observer was trained and then practiced categorizing tape recordings of actual classroom sessions until a Scott coefficient of intra observer consistency of .99 was attained. The observer then observed three classroom situations with different trained observers present during these visits. The interobserver reliability coefficients were .90, .87, and .92 for the three simultaneous observations.

The results showed that first, second, and sixth grade teachers did more talking in social studies than in other subject areas. Third, fourth, and fifth grade teachers did more talking in arithmetic. Student talk was lowest in grade one and two and highest in grades three, four, and five in social studies.

Medley and Mitzel (1958a) developed and applied an ANOVA technique to the reliability of observational data. The model assumed that N teachers were visited n times each by a team of n observers. The assumption of linearity of variances was explicitly stated by the equation

$$X_{ijk} = T_i + V_j + I_{ij} + e_{ijk}$$

in which all the variables actually represented deviations from their respective means. The variables in the equation were defined as follows: X_{ijk} was the deviation from the mean of all values assigned to teacher i during visit j by observer k. T_i was the deviation from the mean of all observations associated with teacher i, V_j the deviation associated with visit j, I_{ij} the deviation of the interaction between teachers and visits, and e_{ijk} the deviation of the residual for teacher i on visit j in observation k. Based on these definitions, I_{ij} was viewed as *visit* error for teacher i on visit j and e_{ijk} as residual or *observer* error. Error was therefore considered to have two components. The first was due to a lack of stability of teacher performance and the "observer" error resulted from the discrepancy between two records of the same teacher performance made by two observers.

The above equation permitted the taking of mathematical expectations and yielded

$$\sigma_x^2 = \sigma_t^2 + \sigma_v^2 + \sigma_{tv}^2 + \sigma^2$$

where σ_x^2 was the total variance for all the observations x, σ_t^2 was the variance of the T_i, σ_v^2 of the V_j, σ_{tv}^2 of the I_{ij}, and σ^2 of the e_{ijk}.

Based on the above, a reliability coefficient based on a single observation was given as

$$R = \sigma_t^2 \,/\, (\sigma_t^2 + \sigma_{tv}^2 + \sigma^2)$$

where the numerator on the right, σ_t^2, was the "true score" variance. This meant that the true score of interest was T_i, the mean of all performances of teacher i on all occasions j on which a visit was possible. The authors (Medley & Mitzel, 1958a) indicated that the σ_v^2 variance component was removed because they compared teachers who had been visited equally often. Since the scores were means over all visits the visit effects cancelled out.

A second reliability coefficient, R', was defined as

$$R' = (\sigma_t^2 + \sigma_{tv}^2) \,/\, (\sigma_t^2 + \sigma_{tv}^2 + \sigma^2)$$

in which the true score was considered to be the performance of teacher i on visit j. This coefficient was actually equivalent to a *coefficient of observer agreement* which usually is calculated as a Pearson r. Here, the fluctuations in teacher performance were considered part of "true score" variance because they were observable by all observers present on a particular occasion.

The reliability of the mean of a number of scores assigned to the same teacher, R_{mn}, was defined in terms of observer team size n and number of visits m

$$R_{mn} = mn\sigma_t^2 / (mn\sigma_t^2 + n\sigma_{tv}^2 + \sigma^2)$$

Estimates of R, R', and R_{mn} were made from an ANOVA which was based on the assumptions that T_i, V_j, I_{ij}, and e_{ijk} were normally and independently distributed in repeated random sampling with zero means and with variances of σ_t^2, σ_v^2, σ_{tv}^2, and σ^2, respectively. The ANOVA is reproduced in Table 1.

Table 1

Medley and Mitzel Reliability ANOVA

		Mean Squares	
Source of Variation	d.f.	Observed	Expected
Teachers	N-1	s_t^2	$\sigma^2 + n\sigma_{tv}^2 + Mn\sigma_t^2$
Visits	m-1	s_v^2	$\sigma^2 + n\sigma_{tv}^2 + Nn\sigma_v^2$
Visit Error	(N-1)(m-1)	s_{tv}^2	$\sigma^2 + n\sigma_{tv}^2$
Observer Error	Nm(n-1)	s^2	σ^2

The ANOVA technique was then applied to the Cornell and Withall techniques. In the first instance, six observers visited 33 teachers in teams of two such that each observer visited each teacher once. In the second, two observers visited four teachers eight times.

The modified Cornell schedule contained the following eight scales: activity, variety, pupil climate, teacher climate, social organization, differentiation, pupil initiative, and content. Both the reliability coefficient, R, and the coefficient of observer agreement, R', were calculated for each scale and were: .41 and .63, .42 and .42, .00 and .00, .32 and .32, .37 and .66, .35 and .64, .00 and .43, and .00 and .23, respectively. The modified Withall categories were: learner-supportive, problem-structuring, neutral, directive, reproving and climate index. The reliability and observer agreement coefficients were .25 and .90, .50 and .98, .00 and .50, .50 and .97, .00 and .88, .47 and .96, respectively.

In their later paper (Medley & Mitzel, 1963) in which the measurement of classroom behavior by systematic observation was discussed, the ANOVA technique for measuring the reliability of observations was further elaborated. This more complete analysis is given in Table 2 and assumed that scores were available for class *c* on *i* items recorded by *r* observers on *s* visits or situations. A typical score was therefore indicated as X_{cris}.

The adaption of this ANOVA to a specific instance was dependent on three rules. The first was to substitute specific numerical values for literal ones, drop any line with zero degrees of freedom, and change the last remaining line to "residual." The second was to omit from the expected mean squares in the remaining lines all the components whose line had been dropped and also the component that corresponded with the new "residual." The third rule was to omit any component in any of the remaining lines that contained a subscript of a "fixed" variable. This rule applied to all but the first component on any line which was never omitted. A "fixed" variable was a variable without an infinite number of values in the population.

The calculation of the reliability of observational data was based on the standard definition

$$\underline{\text{Rho}} = \sigma_T^2 / \sigma_X^2$$

Based on q recorders, j items, and t situations (referred to earlier as r, i, s, respectively) the variance of the population of the true scores was defined as

$$\sigma_T^2 = (qjt)^2 \sigma_c^2$$

where X_{cqjt} was the "true score" and σ_c^2 was the first component shown in Table 2. The general expression for the variance of the actual scores of all the teachers in population about their own mean, σ_X^2, was defined as

$$\sigma_X^2 = qjt(qjt\sigma_c^2+jt\sigma_r^2+qt\sigma_i^2+qj\sigma_s^2+jt\sigma_{cr}^2+qt\sigma_{ci}^2$$
$$+qj\sigma_{cs}^2+t\sigma_{ri}^2+j\sigma_{rs}^2+q\sigma_{is}^2+t\sigma_{cri}^2+j\sigma_{crs}^2+q\sigma_{cis}^2$$
$$+\sigma_{ris}^2+\sigma^2).$$

Table 2

Medley and Mitzel Expanded Reliability ANOVA

Source of Variation	d.f.	Obtained Mean Square	Expected Mean Square
1. Class (C)	c-1	s_c^2	$ris\sigma_c^2+is\sigma_{cr}^2+rs\sigma_{ci}^2+ri\sigma_{cs}^2+s\sigma_{cri}^2+i\sigma_{crs}^2+r\sigma_{cis}^2+\sigma^2$
2. Recorder (R)	r-1	s_r^2	$cis\sigma_r^2+is\sigma_{cr}^2+cs\sigma_{ri}^2+ci\sigma_{rs}^2+s\sigma_{cri}^2+i\sigma_{crs}^2+c\sigma_{ris}^2+\sigma^2$
3. Item (I)	i-1	s_i^2	$crs\sigma_i^2+rs\sigma_{ci}^2+cs\sigma_{ri}^2+cr\sigma_{is}^2+s\sigma_{cri}^2+r\sigma_{cis}^2+c\sigma_{ris}^2+\sigma^2$
4. Situation (S)	s-1	s_s^2	$cri\sigma_s^2+ri\sigma_{cs}^2+ci\sigma_{rs}^2+cr\sigma_{is}^2+i\sigma_{crs}^2+r\sigma_{cis}^2+c\sigma_{ris}^2+\sigma^2$
5. C X R	(c-1)(r-1)	s_{cr}^2	$is\sigma_{cr}^2+s\sigma_{cri}^2+r\sigma_{cis}^2+\sigma^2$
6. C X I	(c-1)(i-1)	s_{ci}^2	$rs\sigma_{ci}^2+s\sigma_{cri}^2+r\sigma_{cis}^2+\sigma^2$
7. C X S	(c-1)(s-1)	s_{cs}^2	$ri\sigma_{cs}^2+i\sigma_{crs}^2+r\sigma_{cis}^2+\sigma^2$
8. R X I	(r-1)(i-1)	s_{ri}^2	$cs\sigma_{ri}^2+s\sigma_{cri}^2+c\sigma_{ris}^2+\sigma^2$
9. R X S	(r-1)(s-1)	s_{rs}^2	$ci\sigma_{rs}^2+i\sigma_{crs}^2+c\sigma_{ris}^2+\sigma^2$
10. I X S	(i-1)(s-1)	s_{is}^2	$cr\sigma_{is}^2+r\sigma_{cis}^2+c\sigma_{ris}^2+\sigma^2$
11. C X R X I	(c-1)(r-1)(i-1)	s_{cri}^2	$s\sigma_{cri}^2+\sigma^2$
12. C X R X S	(c-1)(r-1)(s-1)	s_{crs}^2	$i\sigma_{crs}^2+\sigma^2$
13. C X I X S	(c-1)(i-1)(s-1)	s_{cis}^2	$r\sigma_{cis}^2+\sigma^2$
14. R X I X S	(r-1)(i-1)(s-1)	s_{ris}^2	$c\sigma_{ris}^2+\sigma^2$
15. Residual	(c-1)(r-1)(i-1)(s-1)	s^2	σ^2

The rule for adapting this expression to a particular instance was to drop any component whose subscripts remained constant in all obtained scores. For example, if the same items were used in all classes $qt\sigma_i^2$ would be dropped from the equation defining σ_x^2.

Once σ_T^2 and σ_X^2 have been defined, linear equations were obtained by setting the actual mean squares, which were unbiased estimates of the expected mean squares, equal to their respective expected mean squares. The set of linear equations thus obtained was solved and yielded values which estimated the parameter values from the sample values. The results were as given in Table 3.

The model was then applied to data available on "pupil interest" scores from OScAR3F. The data were collected by two observers in five situations in 24 classes. The given application first considered *items* and *situations* finite and *recorders* infinite, and then *items* finite and *situations* and *recorders* infinite.

It was pointed out that the proposed reliability calculation did not require the assumption of normality of the distribution. This was so because the expected mean squares, upon which the reliability calculation depended, did not require that one assume a normal distribution. However, often one wished to test hypotheses regarding the value of the components for which the assumption of normality was required so that *F* tests could be made.

Denny (1968) reported on the reliability and validity of the Denny Rusch, Ives Classroom Creativity Observation Schedule. The schedule was constructed to identify those teacher-pupil variables which were related to pupil gain on creativity measures and contained three dimensions: climate, general structuring, and specific structuring. There were 11 items comprising the schedule. These were: motivational climate, pupil interest, teacher-pupil relationship, and pupil-pupil relationship-climate, pupil initiative, teacher approach, and adaption to individual differences, variation in materials and activities--general structuring, encouragement of pupil divergent thinking, encouragement of unusual pupil responses, and uniqueness--specific structuring.

Thirty sixth grade classes in a Midwestern state were visited three times by trained observer teams of three recorders. The observations were made between pre and post-tests on adaptations of Guilford's tests.

Both of the Medley and Mitzel (1958a, 1963) techniques

Table 3

Medley and Mitzel Estimation of Variance Components

---

1. σ_c^2 (=)[a] $\frac{1}{ris}(s_c^2-s_{cr}^2-s_{ci}^2-s_{cs}^2+s_{cri}^2+s_{cis}^2+s_{crs}^2-s^2)$

2. σ_r^2 (=) $\frac{1}{cis}(s_r^2-s_{cr}^2-s_{ri}^2-s_{rs}^2+s_{cri}^2+s_{ris}^2+s_{crs}^2-s^2)$

3. σ_i^2 (=) $\frac{1}{crs}(s_i^2-s_{ci}^2-s_{ri}^2-s_{is}^2+s_{cri}^2+s_{cis}^2+s_{ris}^2-s^2)$

4. σ_s^2 (=) $\frac{1}{cri}(s_c^2-s_{cs}^2-s_{rs}^2-s_{is}^2+s_{crs}^2+s_{cis}^2+s_{ris}^2-s^2)$

5. σ_{cr}^2 (=) $\frac{1}{is}(s_{cr}^2-s_{cri}^2-s_{crs}^2+s^2)$

6. σ_{ci}^2 (=) $\frac{1}{rs}(s_{ci}^2-s_{cri}^2-s_{cis}^2+s^2)$

7. σ_{cs}^2 (=) $\frac{1}{ri}(s_{cs}^2-s_{crs}^2-s_{cis}^2+s^2)$

8. σ_{ri}^2 (=) $\frac{1}{cs}(s_{ri}^2-s_{cri}^2-s_{ris}^2+s^2)$

9. σ_{rs}^2 (=) $\frac{1}{ci}(s_{rs}^2-s_{crs}^2-s_{ris}^2+s^2)$

10. σ_{is}^2 (=) $\frac{1}{cr}(s_{is}^2-s_{cis}^2-s_{ris}^2+s^2)$

11. σ_{cri}^2 (=) $\frac{1}{s}(s_{cri}^2-s^2)$

12. σ_{crs}^2 (=) $\frac{1}{i}(s_{crs}^2-s^2)$

13. σ_{cis}^2 (=) $\frac{1}{r}(s_{cis}^2-s^2)$

14. σ_{ris}^2 (=) $\frac{1}{c}(s_{ris}^2-s^2)$

15. σ^2 (=) s^2

---

[a] The symbol (=) is to be read "is estimated by."

were used. The latter model to estimate the total reliability of .42, and the former to calculate R, R', and R_{mn} for each of the 11 items comprising the schedule. The values of R ranged from .15 to .72, of R' from .40 to 1.00, and of R_{mn} from .38 to .91. The author (Denny, 1968) concluded that the reliability estimates were at least as good as those obtained for similar schedules but that the validity estimates indicated a need for further analysis of the dimensions and items of the schedule.

Medley (1967) described a new way to score the OScAR 4V so that more meaningful information could be obtained from the raw scores. The method depended on an ANOVA technique which permitted the partitioning of variance through the use of orthogonal contrasts. The data were collected by an observer who visited 70 teachers four times for about 20 minutes per time. These scores were correlated with scores collected a week or two later on four more visits. The correlations were estimated for each scale by the 1958 ANOVA technique for the four "entry" and six "exit" categories. The entry categories were: pupil initiative, cohesion, divergence and a total score. The exit categories were: feedback, valence enthusiasm, positivity, encouragement, and a total score. The intercorrelation between the total scores was .73 and the range for the other intercorrelations was 0 to .78.

The rationale for the use of orthogonal contrasts to develop scoring keys was that the transformed scores remained linear independent functions of the original raw scores while at the same time the contrasts yielded information of specific interest to the investigator. This was so because the contrasts could be chosen in a very large number of ways. It was therefore incumbent on the investigator to choose the set which best answered whatever questions interested him most.

Brown et al.(1968) calculated "between observer," "within-observer" and internal consistency reliability as well as the correlations mentioned earlier in the preceding section of this chapter (see p.10). The authors (Brown et al., 1968) pointed out that since they were using "untrained" observers the Medley and Mitzel (1963) ANOVA technique was not suitable because the ANOVA gave a reliability measure of "between observer" variability while what was needed for their data was a "within observer" reliability coefficient. Accordingly, a within observer reliability coefficient for the repeated viewing of the films was derived. The within observer reliabilities ranged from .48 to .62.

The derivation was based on the rationale that if the same observer scored the same teaching situation twice in the same way, then the judge's scoring was reliable. To accomplish this end, the ratio of two different values for the variance of the difference between the first and second viewings was derived and constituted the reliability coefficient.

The difference d_i, was defined as

$$d_i = x_{1i} - x_{2i}$$

where 1 and 2 refer to the viewing and the i refers to the items.

Then, for independent scores

$$\begin{aligned} V(d_i) &= V(X_{1i} - X_{2i}) \\ &= V(X_{1i}) + V(X_{2i}) \\ &= \sigma^2 + \sigma^2 = 2\sigma^2 \end{aligned}$$

However, for correlated scores

$$\begin{aligned} V(d_i) &= V(X_{1i} - X_{2i}) \\ &= V(X_{1i}) + V(X_{2i}) - 2\ \mathrm{Cov}\ (X_{1i}, X_{2i}) \\ &= 2\sigma^2 - 2\sigma_{12} = \sigma_d^2 \end{aligned}$$

These two formulas were based on the assumptions that

$$(1)\ V(X_{ij}) = \sigma^2 \text{ for } i - 1,2$$

$$j = 1, \ldots, n$$

and that $$(2)\ p(x) = \frac{1}{k}$$

where the probability of selecting a particular item, p(x), was equal for all the possible choices, k. The reliability coefficient was defined as

$$\underline{Rho}_{jf} = 1 - \sigma_d^2 \ / \ 2\sigma^2$$

The value of σ^2 was treated as a constant because of the assumption of random choice of each item x on the part of the judge and was calculated as

$$\sigma^2 = \underset{x}{\mathrm{SUM}}(x-u)^2 p(x)$$

The actual reliability calculation was then given as

$$r_{jf} = 1 - s_d^2 / 2\sigma^2$$

where s_d^2, the sample value, estimated σ_d^2.

Rho and its statistic, r_{jf}, were equal to the difference between perfect correlation and the ratio of the difference between the variance for independent scores and the covariance of dependent scores to the variance of independent scores. As a result, for independent scores Rho becomes equal to zero. This was exactly the formulation that Brown et al. (1968) wanted because they were interested in a measure of agreement within the observers. The greater the agreement the greater the value of Rho and r_{jf}. Mathematically, this can be seen easily because it can be shown that

$$\underline{Rho} = \sigma_{12} / \sigma^2$$

and therefore Rho is directly proportional to the covariance.

Item reliability was calculated through Kuder-Richardson formulas and ranged from .77 to .81. "Between observer" reliability was reported as fair.

Seibel (1967) investigated whether it was possible to predict the classroom behavior of teachers. The Ss were 100 graduate students with liberal arts backgrounds who were enrolled in the Harvard Graduate School of Education in 1954. The Ss were rated by the classroom teacher in whose room they had their teaching practicum and by their university supervisor on eight criteria of teacher behavior. The criteria were: rewards, support, contact, movement, service, compliance, suggestions, and humor. The ratings of the Ss were adjusted for "reliability" by asking each rater to indicate the "amount of confidence" he had in each of his ratings. Confidence in ratings was indicated on a seven point scale from "complete confidence that rating is accurate"--7, to "no confidence whatever, just a guess"--1. The estimates of confidence were treated as estimates of reliability according to the following scale:

Confidence Rating	Reliability Estimate
7	1.00
6	0.83
5	0.67
4	0.50
3	0.33
2	0.17
1	0.00

The reliability estimates were then used to adjust the behavior ratings according to the formula:

$$x' = rx + (\bar{x} - r\bar{x})$$

where x' = the estimated true rating
r = the estimated reliability of the obtained rating
$\underline{x}$ = the obtained rating
$\bar{x}$ = the mean of the obtained ratings for the group

The eight behavior rating scores were then correlated with the 12 predictor variables which were: Miller Analogies Test, Minnesota Teacher Attitude Inventory score (MTAI), F-scale, Minnesota Multiphasic Personality Inventory Paranoia, Psychastenia, and Social Introversion--Extroversion Scales, Wickman Schedule "no consequence" and extremely grave consequence" Pupil Misbehaviors, Previous Teaching-Leadership Activities with children, Practice Teaching Grade, change in MTAI, change in F-scale.

Zero order, multiple, and canonical correlations were found and led Seibel (1967) to conclude that ". . . to a degree it may be possible to predict how a teacher will behave in the classroom."

Summary of Literature on Recent Methods of Calculating Reliability of Observational Data

The papers reviewed in this section were different from those in the review of traditional methods in that they were more involved with the problem of the reliability of observational data than the studies reviewed in the previous section. The most heuristic and technically advanced method of calculating reliabilities was that of Medley and Mitzel (1958a, 1963).

Of the other work presented, one study developed a reliability coefficient which was actually a percentage of agreement (Scott, 1955) while another developed a reliability estimate based on confidence ratings (Seibel, 1967). Two other studies adapted or used the Scott coefficient (Flanders, 1960; Furst & Amidon, 1967), while two studies used the ANOVA techniques of Medley and Mitzel (Denny, 1968; Brown et al., 1968). Of these latter two studies, Denny actually used both ANOVA models without any adaptation or change. Brown et al. also used the ANOVA technique but at the same time developed their own "within-observer" reliability coefficient.

Summary of Related Literature

The research in this chapter was categorized under two headings which dealt with traditional and other methods of estimating reliabilities of observational data. The usual method of calculating reliabilities was found to be the Pearsonian correlation or the percentage of agreement or their equivalents. Only two studies other than those of Medley and Mitzel used an ANOVA technique.

The difficulty in applying the factorial model, besides its theoretical complexity, was due to its administrative difficulties. These difficulties resulted from the requirement that the same observers visit the same teachers more than once. The present investigation sought to develop procedures which would permit reliability estimates under the more typical field situations.

Chapter II
The Subjects, Materials, and Procedures

The purpose of this investigation was to study the relationship of the variables which were present during observations that were carried out by members of a team. This information was to be applied so that the reliability of this type of data could be estimated through an ANOVA technique or techniques under different conditions. The model was then to be applied to the SUTEC Observation Schedule.

The aims of this section were: (1) to describe the subjects of an observational study in general, and the subjects who participated in the SUTEC study in particular; (2) to describe the materials; (3) to indicate the procedures which were followed; and (4) to present the statistical bases for the analyses of the data.

The Subjects

A study dealing with observational data usually has two different sets of subjects, those being observed and those doing the observing. At different stages of the investigation, one is interested in first one of these sets of subjects and then the other. For the purposes of this study, those being observed were the teachers and those doing the observing were the people who constituted the observer team.

At the beginning of an observational study the major problems are those which pertain to the observers and their ability to see and report accurately that which they have been instructed to observe. This investigation addressed itself to this question of the reliability of observations and therefore the subjects under consideration were the observers.

The training and employment of a team of observers is usually expensive. For this reason, observer teams are generally restricted to 10 or fewer members. The number of teachers visited by the team, when reliability is to be established, is also generally less than 10. The maximum number of teachers actually visited, as found in the review of the literature in the previous chapter of this paper, was six.

There were 10 people who were trained and acted as observers for the SUTEC project. All of the observers were graduate students in education, related areas, or their equivalent. The five teachers who were observed by the team, for the reliability study, were all regularly licensed New York City teachers who had a minimum of three years of teaching experience.

The Materials

The general model presented in this investigation is applicable to many different types of observation schedules. The types of materials involved fall under the generic definition given in the first chapter of this report. For examples of schedules to which the model is applicable, the reader is referred to Chapter I.

In terms of the SUTEC data, the observer team observed only the following seven categories of behavior: teacher mobility, involvement of children, materials present, materials in use, directed behavior, spontaneous behavior, and irrelevant acts. These items are briefly described below. The underlying rationale of the schedule and more detailed descriptions were given by Chapline (1968). A copy of the schedule is attached (Appendix).

Teacher Mobility. The number of different positions occupied by the teacher during the second five minutes of each learning activity--indicated on a room sketch.

Involvement of Children. A global judgment of the attentiveness of the whole class during each learning activity--assessed on a three point scale from uninvolved (1) to highly involved (3).

Materials Present. The number of different materials present during the entire observation--checked on a list of materials.

Materials in Use. The number of different materials in use during the entire observation--checked on a list of materials.

Directed Behavior. The number of times during each activity that the teacher called on pupils without the pupils first indicating a willingness to respond.

Spontaneous Behavior. The number of times that the pupils indicated a willingness to respond before being asked to do so, plus the number of times that the pupils responded spontaneously before permission was granted. The score on this category was weighted in a ratio of 1:2, respectively, before being added. Raising hand behavior would be scored as a one while calling out the answer would be scored as a two. If both occurred during the same activity, the activity would be scored as a three provided nothing else happened for the duration of the activity.

Irrelevant Acts. The number of acts or movements obviously not related to the learning activity of twelve randomly selected children.

The schedule yielded raw scores on the seven categories. Involvement of children had a range from one to three, while the other six categories had no specific range built into the schedule and were actually frequency counts.

The Procedures

The initial step in the procedure was the estimation of the reliability of three of the categories that were considered for inclusion in the final form of the SUTEC schedule. These items were mobility, involvement, and irrelevant acts. For this reliability estimate, seven observers visited two teachers and rated them on three categories. Only four of the observers who made the second visit were also present during the first visit.

The next step was to estimate the reliability of the entire schedule. Three other teachers were visited by an observer team of seven members. Although careful planning had preceded the visits to insure that all 10 members of the observation team would be present, such was not the case. Here, too, the seven observers present were not the same in each case.

The ten members of the observer team were each given a copy of the observation schedule they were to use and the categories were discussed and explained to their satisfaction. This discussion was followed by a field test which, in turn, was followed by a comparison and discussion of the obtained results. Upon repetition of this procedure, the observer team felt confident in their ability to use the schedule properly. Visits to the different teachers by the entire group were arranged to determine the reliability of the observer team. All the observations were conducted through a one way mirror with the teacher's knowledge and consent.

As was evident from Chapter I, the method of training the SUTEC team was consonant with generally accepted practice.

The Statistical Procedures

This section dealt with some of the theoretical considerations that pertained to the problem. The areas discussed were: some of the shortcomings of the traditional methods of calculating reliabilities, the meaning of "crossed" and "nested" factors, the relationship of ANOVA to reliability estimation, the variables and some of the conditions under which different designs are possible, the calculation of expected mean squares, and some of the general models under the various conditions.

Difficulties with Traditional Reliability Estimates. The shortcomings of the product moment coefficient of correlation and the percent of agreement between observers as measures of reliability of observational data were originally pointed out by Medley and Mitzel (1958a, 1963) and paraphrased by Brown et al. (1968). For one thing, the sampling distribution of r is dependent on N, the number of scores on each item, and it is difficult to have large numbers of people view the same classroom on two different occasions or to control variations between the two visits. Furthermore, the number of classrooms visited by two different observers, at two different times is likely to be small. In either case an N as great as 100 in dealing with observational studies is extremely rare. With N = 100 the confidence interval for the correlation coefficient may be as wide as .33 (Medley & Mitzel, 1963) and therefore the correlation coefficient is not very precise. At the same time most such correlations are usually based on total scores which do not take into account variations in scoring individual items or categories.

Percentage of agreement between observers may give very little information about the reliability of scores obtained. This is possible if the observed teaching practice occurs in each room. For then, the reliability of that item as a differentiator of teachers will be zero. It is equally posiible that near perfect agreement be reached about the number of times that a teacher employed a certain category of behavior, and if the teacher sharply reversed these behaviors from observation to observation the reliability of these categories from visit to visit would be zero.

The shortcomings mentioned above led Medley and Mitzel (1963) to develop their single intraclass correlation coefficient. The estimate of Rho so obtained was more precise than any combination of interclass correlations because such a combination of correlation coefficients was not made up of independent measures. The Medley and Mitzel (1963) model permitted the calculation of the variance attributable to each of the independent factors operating during the course of the observations. At the same time, the different reliability coefficients appropriate to the uses to which the scores might be put could all be estimated from the one analysis of variance.

Crossed and Nested Factors. Factors are said to be *crossed* if each level of each factor appears at least once with every level of every factor. Factors are said to be *nested* if each level of each factor appears in only one *level* of the other factors (Millman & Glass, 1967). A factor which is *not* nested in any other factor may therefore be considered *crossed* (Peng, 1967).

Relationship of ANOVA to Reliability Estimation. The definition of reliability given by Medley and Mitzel (1963) was comparable to that discussed by Winer (1962) for a more simplified type of design. However, because of the comparability of the concepts parts of the argument will be reproduced and some of the algebra that was deleted will be filled in. The basic definition, that the reliability of k measurements is the ratio of the variance of the true scores to the sum of the true score variance and the variance due to errors of measurement, was the same for both authors.

Winer (1962) indicated that the reliability for the mean of k measurements may be estimated by

$$r_k = \frac{(1/k)(MS_{\text{between people}} - MS_{\text{w. people}})}{(1/k)(MS_{\text{between people}} - MS_{\text{w. people}}) + (1/k)MS_{\text{w.people}}}$$

where the variance of the true score, σ_t^2, was estimated by the numerator and the sum of the true score and error of measurement variances, σ_x^2, was estimated by the denominator of r_k. Multiplication of both numerator and denominator of r_k by k yielded,

$$r_k = \frac{MS_{\text{between people}} - MS_{\text{w. people}}}{(MS_{\text{between people}} - MS_{\text{w. people}}) + MS_{\text{w. people}}}$$

$$= \frac{MS_{\text{between people}} - MS_{\text{w. people}}}{MS_{\text{between people}}}$$

$$= \frac{MS_{\text{between people}}}{MS_{\text{between people}}} - \frac{MS_{\text{w. people}}}{MS_{\text{between people}}}$$

$$= 1 - \frac{MS_{\text{w. people}}}{MS_{\text{between people}}}$$

The estimate of σ_t^2 was therefore seen to be $MS_{\text{between people}} - MS_{\text{w. people}}$, the difference between the sources of variation while the estimate for σ_x^2 was the sum of σ_t^2 and the error of measurement.

It was also shown that r_1 the reliability estimate of a single measurement was given by

$$r_1 = \frac{(1/k)\,(MS_{\text{between people}} - MS_{\text{w. people}})}{(1/k)\,(MS_{\text{between people}} - MS_{\text{w. people}}) + MS_{\text{w. people}}}$$

Multiplying both numerator and denominator by k yielded

$$r_1 = \frac{MS_{\text{between people}} - MS_{\text{w. people}}}{MS_{\text{between people}} + (k-1)MS_{\text{w. people}}}$$

For the sake of algebraic brevity the following substitutions were made.

$X = (1/k)\,(MS_{\text{between people}} - MS_{\text{w. people}})$

$Y = MS_{\text{w. people}}$

Upon substitution into the equation which defined r_1

$r_1 = X / (X+Y)$

multiplying, $r_1\,(X+Y) = X,$

dividing by r_1 $X + Y = X / r_1$

transposing, $Y = \frac{X}{r_1} - X,$

combining, $Y = (X - r_1X) / r_1$

Substituting back into the original equation which defined r_k.

$$r_k = \frac{X}{X + \frac{1}{k}\left[(X - r_1X) / r_1\right]},$$

multiplying numerator and denominator by kr_1,

$$r_k = kr_1X / \left[kr_1X + (X - r_1X)\right]$$

factoring,

$$r_k = kr_1X / \left[X\,(kr_1 + 1 - r_1)\right]$$

$$r_k = kr_1 / \left[1 + r_1\,(k-1)\right]$$

This formula is the well known Spearman-Brown prediction formula. A somewhat different treatment which yielded the same result was given by McNemar (1962).

The simplified assumptions upon which these formulas and calculations rest are that $MS_{\text{w. people}}$ may be pooled to provide an estimate of the error of measurement, that the error of estimate is uncorrelated with the true score, and that the sample of n people and k measuring instruments are random samples to and from which generalizations are to be made, respectively. The more involved cases when these assumptions were not met need not be discussed here because the essential relationship has been indicated and the calculation of σ_t^2 and σ_x^2 can be estimated by following the "rules of thumb" given by Medley and Mitzel (1963).

The Variables and Designs

Before considering specific designs, some basic notions about the composition of a specific score and its relationship to population parameters and "error" will be discussed. The discussion will be presented in terms of a simplified case and will be alluded to once again in Chapter III of this report.

The factors which may be expected to produce variation among observational scores of teacher behaviors are differences among teachers and differences among visits. For convenience, T_i will be used to represent the difference between the mean for teacher i and the mean of all the observations, and V_j will represent the difference between the mean of visit j and the mean of all the visits. In essence then, T_i and V_j represent the deviations associated with teacher i and visit j. It is assumed that T_i and V_j will be the same for teacher i on every visit and for all teachers on the jth visit to each of them, respectively.

Human behavior being what it is, it is probable that some teachers will behave differently on the first visit than on the others, while other teachers' behaviors may change slightly from visit to visit so that there may be a great change in the behavior of some of the teachers from the initial to the final visit. In statistical terms, one would say that there is an interaction between visits and teachers. If P_{ij} denotes the performance of teacher i on visit j and I_{ij} the interaction term, then

$$P_{ij} = u + T_i + V_j + I_{ij}$$

where u denotes the grand mean of all teachers on all visits. The score, X_{ijk}, assigned to teacher i by observer k for visit j may or may not be identical to the actual performance P_{ij} of that teacher on that visit. The "error" is defined as the difference between the assigned score and the actual performance.

$$e_{ijk} = X_{ijk} - P_{ij}$$

Substituting for P_{ij},

$$e_{ijk} = X_{ijk} - (u + T_i + V_j + I_{ij}).$$

Transposing, $X_{ijk} = u + T_i + V_j + I_{ij} + e_{ijk}$

This simplified model follows closely that presented by Medley and Mitzel (1958a) and in principle is easily generalized to include other factors. This linear model leads to

$$\sigma_x^2 = \sigma_t^2 + \sigma_v^2 + \sigma_{tv}^2 + \sigma^2$$

where σ_x^2 is the total variance for all observations X, σ_t^2 is the variance of T_i, σ_v^2 of the V_j, σ_{tv}^2 of the I_{ij}, σ^2 of the e_{ijk}.

The variables which were considered essential for classroom observations were those used by Medley and Mitzel (1963). These were: teacher or classes, observers, items, and situations. Under certain conditions one or more of the variables may be deleted from the analysis. For example, if all the teachers were rated on only one item by the team of observers, or if only one observer did all the observations, or if each teacher is visited only once, then items, observers, and situations would have to be dropped from the analysis associated with their respective cases.

Within each of the above contingencies it is possible to have more than one condition operating concurrently.

An example that will be considered in some detail in the next chapter, is the case of the partially nested or partially hierarchical design. If each teacher were visited only once by a team of observers which had the same number of people on the team but not the same team, the team factor would be nested under the teacher factor. To consider a simple case, suppose that two teachers were visited by three observers and rated on four items. Each teacher essentially has a team peculiar to himself because the people comprising the team for teacher one are not all in the team for teacher two, etc. A schematic representation of this situation is given in Table 4.

Table 4

Schematic Representation of a Three Factor Partially Nested Design

	Teacher 1			Teacher 2		
	Obs. 1	Obs. 2	Obs. 3	Obs. 4	Obs. 5	Obs. 6
Item 1						
Item 2						
Item 3						
Item 4						

The general model for this design has the structural form (Winer, 1962):

$$ABC_{ijk} = u + A_{i} + B_{j(i)} + C_{k} + AC_{ik} + BC_{j(i)k} + \text{error}$$

where teachers, observers, and items are factors A, B, and C, respectively, and the terms on the right side of the equation represent population parameters. In this design it is possible that (1) each observation schedule item has n subcategories and therefore each cell has n scores or that (2) each item has no subcategories and yield only one score per cell. In the former case the left hand side of the equation and the error term represent the mean of the measurements for the n elements under the treatment combination ijk, where i, j, k are the number of elements in factors A, B, C, respectively, and the error term is the average error within the respective cells. In the latter case, there being no within cell variation, the ABCijk and the error term refer to the actual scores and errors attained under treatment condition ijk rather than to the cell values. This model will be used in Chapter III of this investigation where it will be further discussed.

Before continuing with an alternate design for analyzing this type of situation, definitions of "random," "finite," and "fixed" factors will be given. A factor is random if its levels resulted from a random sampling taken from a population of levels with normally distributed effects. "Teachers" is an example of a factor that is usually considered random. If a random sample from a finite population of levels constitutes the factor in a study, the factor is considered finite. When a systematic selection of levels, all levels, or only levels of interest to the investigator are included in a study, the factor is considered fixed. In each case, the results of the ANOVA are generalizable to the population from which the levels were drawn. In terms of the variance components, a random factor must be considered throughout the analysis and is contained as a component in each expected mean square term associated with each "source" of variation. Fixed factors on the other hand are not carried throughout the analysis. The procedures for determining the expression for the expected mean squares are elaborations and applications of this concept (Millman & Glass, 1967). This point is further elaborated in Chapter III.

An alternative approach to the above example in which there was only one observation per cell is to treat the situation as a two factor repeated measure design. Table 5 shows repetitions of the items three times, O_1, O_2, and O_3 and again O_4, O_5, and O_6.

Table 5

Schematic Representation of Two Factor Repeated Measures Design

	Observers	Item 1	Item 2	Item 3	Item 4
Teacher 1	O_1				
	O_2				
	O_3				
Teacher 2	O_4				
	O_5				
	O_6				

The structural model for this design has the following form (Winer, 1962):

$$X_{ijk} = u + A_i + P_{k(i)} + B_j + AB_{ij} + BP_{jk(i)} + \text{error}$$

where teachers and items are factors A and B, respectively, and $P_{k(i)}$ and the error term are the effects of observer k who is nested under a "teacher" level and the error associated with the observer, respectively. Here too, the variables on the right hand side of the equation represent parameters.

Generally, an experiment in which the same elements are exposed to n treatments requires n observations on each element. Hence the term repeated measure. The purpose of this type of experiment, especially in "learning" studies is to provide a control on differences between subjects. This is accomplished because each subject essentially serves as his own control. To the degree that specific characteristics of the individual elements remain constant under different treatment conditions, observations on the same elements tend to be positively correlated or dependent. An alternative approach to a repeated measures design is to include a nested random factor, a dummy variable, in the model to absorb the correlation between the experimental errors. This approach was followed in ChapterIII and was recently discussed as a possible way to adapt existing computer programs to correlated observations (Clifford, 1968).

In a two factor design in which there are repeated measures on factor B the comparisons between the treatments at different levels of factor A involve differences between groups as well as differences associated with factor A. Under these conditions the main effects of factor A are said to be confounded with differences between groups whereas the main effects of factor B and the AB interaction term are free of such confounding. Tests on factors which are not confounded are more sensitive because there are fewer uncontrolled sources of variation.

If one uses the approach that assumes correlated errors, the expected mean square of A has this form

$$E(MS_a) = \sigma^2 \left[1 + (b - 1) r\right] + nb\sigma_a^2$$

where r is the correlation between pairs of observations on the same element and b is the number of levels of factor B. The denominator of an F ratio testing the variance of factor A is equal to the first part of the right hand side of the $E(MS_a)$. *F* tests for the B and AB terms have a denominator whose expected value is $\sigma^2(1-r)$. Therefore, if there is a positive correlation between pairs of measurement the factors which are not confounded have more sensitive tests because their experimental error term is smaller.

In terms of the alternative approach which postulates a nested random factor the $E(MS_a)$ for factor A is:

$$E(MS_a) = \sigma_e^2 + b\sigma_p^2 + nb\sigma_a^2$$

and the denominator of factor A's *F* ratio is equal to the first two terms on the right side of the equation. The denominator for tests on B and AB, the non confounded factors, has the form $\sigma_e^2 + \sigma_{bp}^2$ where σ_{bp}^2 is the subject treatment interaction. The magnitude of σ_{bp}^2 is usually considereably smaller than that of σ_p^2. The above ideas are generalizable to more than two factor designs and specific attention is given to the alternative, nested factor, approach in Chapter III.

Reference to Table 5 indicates that each item is repeated three times for each teacher or that there are three item scores per teacher. This is equivalent to the measuring instrument being applied to each teacher three times. Table 5 to 10 deal with designs in which the *item* factor is repeated.

Upon using the same notation as Medley and Mitzel (1963) where teachers are classes (c), observers are recorders (r), and items are items (i) and considering the last term as the "residual," the design became as shown in Table 6.

In table 6 the D_c, D_r, and D_i, are equal to *zero* or *one* depending on whether the c, r, and i factors *are fixed* or *random*, respectively. This point will be considered again in Chapter III. At the same time the *residual* term would more precisely have been expressed as *Items X Recorders within classes* with degrees of freedom as given and an E(MS) of $D_r\,\sigma_{ri}^2 + \sigma_e^2$. However, just as the factorial model pooled the fourth order interaction, C X R X I X S, with the residual variance, here too, the highest order interaction was pooled with the error term. Therefore, the error term σ_e^2 was replaced throughout by σ^2 which equalled $D_r\sigma_{ri}^2+\sigma_e^2$. For the purist, the full model can be reclaimed by substituting $D_r\sigma_{ri}^2 + \sigma_e^2$ for σ^2 throughout the fourth column of Table 6. However, for the repeated measured designs there is no calculation for

Table 6

Two Way ANOVA for Repeated Measures

Source of Variation	Degrees of Freedom	Obtained Mean Square	Expected Mean Square
Between Recorders	$rc - 1$		
1. Class	$c - 1$	s_c^2	$ri\sigma_c^2 + iD_r\sigma_r^2 +$ $rD_i\sigma_{ci}^2 + \sigma^2$
2. Recorders w. classes	$c(r - 1)$	s_r^2	$iD_r\sigma_r^2 + \sigma^2$
Within Recorders	$rc(i - 1)$		
3. Items	$i - 1$	s_i^2	$rc\sigma_i^2 + rD_c\sigma_{ci}^2 + \sigma^2$
4. C X I	$(c - 1)(i - 1)$	s_{ci}^2	$r\sigma_{ci}^2 + \sigma^2$
5. Residual	$c(r - 1)(i - 1)$	s^2	σ^2

an "error" term and therefore the I X $R_{W:\ \text{classes}}$ term serves as the denominator for the Within Recorders *F* ratios. Therefore, for all intents and purposes the use of $\overline{\sigma}^2$ as the residual *E*(MS) is equivalent to the more cumbersome $D_r\sigma_{ri}^2 + \sigma_e^2$.

To complete the analysis, the linear equation in which the actual MS's serve as estimates for the expected MS's must be solved. Solution of these equations under the assumption of random factors yielded the results indicated in Table 7. The expression on the right of Table 7 makes the variance components estimates specific to the example cited earlier in which c = 2 (2 teachers), r = 6 (6 observers), i = 4 (4 items).

Based on the above and coupled with their (Medley & Mitzel, 1963) "rules of thumb," the definitions of σ_T^2 and σ_X^2 yielded, respectively

$$\sigma_T^2 = (6 \cdot 4 \cdot 1)^2\sigma_c^2 = 576\sigma_c^2$$

Table 7

Estimation of Variance Components for a
Two Factor Repeated Measures Design

1.	$\sigma_c^2 \; (=) \frac{1}{ri} (s_c^2 - s_r^2 - s_{ci}^2 + s^2)$	$(=) \frac{1}{24} (s_c^2 - s_r^2 - s_{ci}^2 + s^2)$
2.	$\sigma_r^2 \; (=) \frac{1}{i} (s_r^2 - s^2)$	$(=) \frac{1}{4} (s_r^2 - s^2)$
3.	$\sigma_i^2 \; (=) \frac{1}{rc} (s_i^2 - s_{ci}^2)$	$(=) \frac{1}{12} (s_i^2 - s_{ci}^2)$
4.	$\sigma_{ci}^2 (=) \frac{1}{r} (s_{ci}^2 - s^2)$	$(=) \frac{1}{3} (s_{ci}^2 - s^2)$
5.	$\sigma^2 \; (=) \quad s^2$	$(=) \quad s^2$

and $$\sigma_X^2 = (6\cdot4\cdot1)\,(6\cdot4\cdot1\ \sigma_r^2 + 4\cdot1\ \sigma_r^2 + 6\cdot1\ \sigma_{ci}^2 + \sigma^2)$$
$$= 24(24\sigma_c^2 + 4\sigma_r^2 + 6\sigma_{ci}^2 + \sigma^2).$$

Therefore, $$r = \sigma_T^2 / \sigma_X^2 = 24\sigma_c^2 / (24\sigma_c^2 + 4\sigma_r^2 + 6\sigma_{ci}^2 + \sigma^2).$$

In the previous design there was only one repeated factor. This was a two factor design because the teachers were visited only once by a three man team of different people for each teacher who rated each item with only one score. Under the same conditions with each item receiving n scores the design may be considered a three factor experiment with repeated measures within the item factor. The measurable variance within each item may change from observer to observer and therefore an interaction component must be added. A schematic representation of this situation is given in Table 8. This may be considered a 2 X 4 X 6 factorial design with repeated measures on the last factor, n = 2. Subscore eight may be represented symbolically as $P_{2(41)}$, that is, the second subscore in G_{41}.

The structural model in which n is the number of subscores for this design may be given as (Winer, 1962)

Table 8

Representation of Three Factor Repeated Measures Design with n Subscores

			Observers					
Teachers	Items	Subscores	O_1	O_2	O_3	O_4	O_5	O_6
T_1	Item 1	1	G_{11}	G_{11}	G_{11}	G_{11}	G_{11}	G_{11}
		2	G_{12}	G_{12}	G_{12}	G_{12}	G_{12}	G_{12}
	Item 2	1	G_{21}	G_{21}	G_{21}	G_{21}	...	
		2	G_{22}	G_{22}	G_{22}	...		
	Item 3	1	G_{31}	G_{31}	...			
		2	G_{32}	...				
	Item 4	1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		2						
T_2	Item 1	1						
		2						
	Item 2	1						
		2						
	Item 3	1						
		2						
	Item 4	1						
		2	G_{42}	G_{42}	G_{42}	G_{42}	G_{42}	G_{42}

$$X_{ijkm} = u + A_i + B_j + AB_{ij} + P_{m(ij)} + C_k + AC_{ik}$$
$$+ BC_{jk} + ABC_{ijk} + CP_{km(ij)} + \text{error}$$

where the right member of the equation contains the parameters A, B, and C which are the teachers, items, and observers, and m(ij) identifies a subscore within group G_{ij}. The ANOVA for this model is given in Table 9 in terms of classes, items, and recorders. The number of subscores, classes, items, and recorders for the general case will be given as n, c, i, and r. For the specific example the values 2,2,4, and 6, respectively, will be substituted.

The residual term for the Within should really have been $\sigma^2_{rp} + \sigma^2_e$. However, as in the case of the two factor repeated measures design, σ^2 was substituted for the last term and this last erm became the "residual." In lines 5 to 8 of Table 9, the original design can be reclaimed by making the substitution $\sigma^2_{rp} + \sigma^2_e$ for σ^2. In lines 1 to 4 of Table 9, the term that σ^2 replaced was really $D_r\sigma^2_{rp} + \sigma^2_e$. However, since the estimation of the variance components rests on the assumption that r, i, and c are random factors, it simplified the design to incorporate this fact in the error term throughout Table 9. If r were not a random factor there would have been two error terms. The error term for the Between subscores would have been $r\sigma^2_p + \sigma^2_e$, since $D_r = 0$ for a fixed factor, and for the Within subscores, $\sigma^2_{rp} + \sigma^2$. This change would not affect any proposed tests of significance because the denominator for the Between F ratio and the first three lines would all be reduced by $D_r\sigma^2_{rp}$. The F ratios for the Within would all contain $\sigma^2_{rp} + \sigma^2_e$ since the σ^2_{rp} was not multiplied by D_r.

Solution of the linear equations when the obtained MS's were used to estimate the expected MS's yielded the results given in Table 10. It was assumed that D_c, D_i, and D_r were equal to one, or that c, i and r were random factors. To complete the reliability calculation,

$$\sigma^2_T = (6 \cdot 4 \cdot 1)^2\sigma^2_c = 576\sigma^2_c$$

and, $\sigma^2_X = (6 \cdot 4 \cdot 1)\,(6 \cdot 4 \cdot 1\sigma^2_c + 6 \cdot 1\sigma^2_{ci} + 6 \cdot 1\sigma^2_p + 4 \cdot 1\sigma^2_r + 4 \cdot 1\sigma^2_{cr}$
$+\sigma^2_{ir} + \sigma^2_{cir} + \sigma^2)$

therefore, $r = 24\sigma^2_c \,/\, (24\sigma^2_c + 6\sigma^2_{ci} + 6\sigma^2_p + 4\sigma^2_r + 4\sigma^2_{cr} + \sigma^2_{ir} + \sigma^2_{cir} + \sigma^2)$

The designs discussed in Tables 6 to 10 dealt with the case of one repeated factor, the item factor. This was

Table 9

Three Factor ANOVA with Repeated Measures on One Factor

Source of Variation	Degrees of Freedom	Obtained Mean Square	Expected Mean Square
Between Subscores	nci-1		
1. Class (C)	c-1	s_c^2	$nir\sigma_c^2+nrD_i\sigma_{ci}^2+r\sigma_p^2+niD_r\sigma_{cr}^2+nD_iD_r\sigma_{cir}^2+\sigma^2$
2. Items (I)	i-1	s_i^2	$ncr\sigma_i^2+nrD_c\sigma_{ci}^2+r\sigma_p^2+ncD_r\sigma_{ir}^2+nD_cD_r\sigma_{cir}^2+\sigma^2$
3. C X I	(c-1)(i-1)	s_{ci}^2	$nr\sigma_{ci}^2+r\sigma_p^2+nD_r\sigma_{cir}^2+\sigma^2$
4. Subscores (P) w. groups	ci(n-1)	s_p^2	$r\sigma_p^2+\sigma^2$
Within Subscores	nci(r-1)		
5. Recorders (R)	r-1	s_r^2	$nci\sigma_r^2+niD_c\sigma_{cr}^2+ncD_i\sigma_{ir}^2+nD_cD_i\sigma_{cir}^2+\sigma^2$
6. C X R	(c-1)(r-1)	s_{cr}^2	$ni\sigma_{cr}^2+nD_i\sigma_{cir}^2+\sigma^2$
7. I X R	(i-1)(r-1)	s_{ir}^2	$nc\sigma_{ir}^2+nD_c\sigma_{cir}^2+\sigma^2$
8. C X I X R	(c-1)(i-1)(r-1)	$-s_{cir}^2$	$n\sigma_{cir}^2+\sigma^2$
9. Residual	ci(n-1)(r-1)	s^2	σ^2

Table 10

Estimation of Variance Components for a Three Factor Design with Repeated Measures on One Factor

1. $\sigma^2_c \; (=) \; \frac{1}{nir}(s^2_c - s^2_{ci} - s^2_{cr} + s^2_{cir}) \quad (=) \; \frac{1}{48}(s^2_c - s^2_{ci} - s^2_{cr} + s^2_{cir})$

2. $\sigma^2_i \; (=) \; \frac{1}{ncr}(s^2_i - s^2_{ci} - s^2_{ir} + s^2_{cir}) \quad (=) \; \frac{1}{24}(s^2_i - s^2_{ci} - s^2_{ir} + s^2_{cir})$

3. $\sigma^2_{ci} \; (=) \; \frac{1}{nr}(s^2_{ci} - s^2_{cir} - s^2_p + s^2) \quad (=) \; \frac{1}{12}(s^2_{ci} - s^2_{cir} - s^2_p + s^2)$

4. $\sigma^2_p \; (=) \; \frac{1}{r}(s^2_p - s^2) \quad (=) \; \frac{1}{6}(s^2_p - s^2)$

5. $\sigma^2_r \; (=) \; \frac{1}{nci}(s^2_r - s^2_{cr} - s^2_{ir} + s^2_{cir}) \quad (=) \; \frac{1}{16}(s^2_r - s^2_{cr} - s^2_{ir} + s^2_{cir})$

6. $\sigma^2_{cr} \; (=) \; \frac{1}{ni}(s^2_{cr} - s^2_{cir}) \quad (=) \; \frac{1}{8}(s^2_{cr} - s^2_{cir})$

7. $\sigma^2_{ir} \; (=) \; \frac{1}{nc}(s^2_{ir} - s^2_{cir}) \quad (=) \; \frac{1}{4}(s^2_{ir} - s^2_{cir})$

8. $\sigma^2_{cir} \; (=) \; \frac{1}{n}(s^2_{cir} - s^2) \quad (=) \; \frac{1}{2}(s^2_{cir} - s^2)$

9. $\sigma^2 \; (=) \; s^2$

---

due to the teachers being visited only once by different teams of observers. If the experience of the researcher with the items of the schedule has indicated that there are no significant sources of variance as a result of treating the items as a repeated factor and breaking up the Within, then the design may be treated as a factorial design with the item factor treated as a regular factor. If this is the case, the three factor repeated measures design in Tables 9 and 10 can be applied to situations in which one of the other factors are repeated. Thus, the design may be used when the same teachers are visited more than once by different observers or when the same observers visit different teachers. In the former situation the teacher factor would be treated as the repeated measure, while in the latter case the observer factor would be the repeated factor. This would merely require replacing the item factor by the teacher or observer factor, respectively.

The next design will consider a three factor design in which there are two repeated measures and will deal with the case of different teachers being visited by the same observers and rated on the same items. In this situation the observer and item factors are the repeated measures across teachers. The structural model for this design may be indicated as

$$\begin{aligned} X_{ijkm} = U + A_i + P_{m(i)} + B_j + AB_{ij} + BP_{jm(i)} + C_k \\ + AC_{ik} + CP_{km(i)} + BC_{jk} + ABC_{ijk} \\ + BCP_{jkm(i)} + \text{error} \end{aligned}$$

Schematically, this situation may be seen in Table 11 where A, B, C, are the teachers, items, and observers, and $P_{m(i)}$ is the sum of the jk observations on subscore m for teacher i.

The ANOVA model is given in Table 12 (Winer, 1962) where the assumption that c, i, and r are random factors has been incorporated into the expected MS's. This assumption permitted the use of the same error or "residual" term throughout. For, it will be noticed in Table 13, that if $D_r = 1$ and $D_i = 1$, i.e., that r and i are random factors, then $\sigma_e^2 + \sigma_{irp}^2$ appears in each line and is equivalent to the error term of Table 12.

Table 11

Schematic Representation of Three Factor Design with Two Repeated Measures and n Subscores

	Subscore	Item 1	...	Item j	Total
		obs. 1...obs. k	...	obs. 1...obs. k	
	1(i)				$P_1(i)$
	.				.
	.				.
	.				.
T	m(i)				$P_m(i)$
	.				.
	.				.
	.				.
	n(i)				$P_n(i)$

Table 12

Three Factor ANOVA with Repeated Measures on Two Factors

Source of Variation	Degrees of Freedom	Obtained Mean Square	Expected Mean Square
Between Subjects	nc-1		
1. Class (C)	c-1	s_c^2	$nir\sigma_c^2+ir\sigma_p^2+nr\sigma_{ci}^2+r\sigma_{ip}^2$ $+ni\sigma_{cr}^2+i\sigma_{rp}^2+n\sigma_{cir}^2+\sigma^2$
2. Subscores (P) w. groups	c(n-1)	s_p^2	$ir\sigma_p^2+r\sigma_{ip}^2+i\sigma_{rp}^2+\sigma^2$
Within Subjects	nc(ir-1)		
3. Items (I)	i-1	s_i^2	$ncr\sigma_i^2+nr\sigma_{ci}^2+r\sigma_{ip}^2+nc\sigma_{ir}^2$ $+n\sigma_{cir}^2+\sigma^2$
4. C X I	(c-1)(i-1)	s_{ci}^2	$nr\sigma_{ci}^2+r\sigma_{ip}^2+n\sigma_{cir}^2+\sigma^2$
5. I X P	c(n-1)(i-1)	s_{ip}^2	$r\sigma_{ip}^2+\sigma^2$
6. Recorders (R)	r-1	s_r^2	$nci\sigma_r^2+ni\sigma_{cr}^2+i\sigma_{rp}^2+nc\sigma_{ir}^2$ $+n\sigma_{cir}^2+\sigma^2$
7. C X R	(c-1)(r-1)	s_{cr}^2	$ni\sigma_{cr}^2+i\sigma_{rp}^2+n\sigma_{cir}^2+\sigma^2$
8. R X P	c(n-1)(r-1)	s_{rp}^2	$i\sigma_{rp}^2+\sigma^2$
9. I X R	(i-1)(r-1)	s_{ir}^2	$nc\sigma_{ir}^2+n\sigma_{cir}^2+\sigma^2$
10. C X I X R	(c-1)(i-1)(r-1)	s_{cir}^2	$n\sigma_{cir}^2+\sigma^2$
11. Residual	c(n-1)(i-1)(r-1)	s^2	σ^2

Table 13

Expected Mean Square of Three Factor Design with Repeated Measures on Two Factors

Source	E(MS)
1. Class (C)	$nir\sigma_c^2+ir\sigma_p^2+nrD_i\sigma_{ci}^2+rD_i\sigma_{ip}^2+niD_r\sigma_{rp}^2+nD_iD_r\sigma_{cir}^2+D_iD_r\sigma_{irp}^2+\sigma_e^2$
2. Subscores (P) w. groups	$ir\sigma_p^2+rD_i\sigma_{rp}^2+iD_r\sigma_{rp}^2+D_iD_r\sigma_{irp}^2+\sigma_e^2$
3. Items (I)	$ncr\sigma_i^2+nrD_i\sigma_{ci}^2+r\sigma_{ip}^2+ncD_r\sigma_{ir}^2+nD_cD_r\sigma_{irp}^2+D_r\sigma_{irp}^2+\sigma_e^2$
4. C X I	$nr\sigma_{ci}^2+r\sigma_{ip}^2+nD_r\sigma_{cir}^2+D_r\sigma_{cir}^2+D_r\sigma_{irp}^2+\sigma_e^2$
5. I X P	$r\sigma_{ip}^2+D_r\sigma_{irp}^2+\sigma_e^2$
6. Recorders (R)	$nci\sigma_r^2+niD_c\sigma_{cr}^2+i\sigma_{rp}^2+ncD_i\sigma_{ir}^2+nD_cD_i\sigma_{cir}^2+D_i\sigma_{irp}^2+\sigma_e^2$
7. C X R	$ni\sigma_{cr}^2+i\sigma_{rp}^2+nD_i\sigma_{cir}^2+D_i\sigma_{irp}^2+\sigma_e^2$
8. R X P	$i\sigma_{rp}^2+D_i\sigma_{irp}^2+\sigma_e^2$
9. I X R	$nc\sigma_{ir}^2+nD_c\sigma_{cir}^2+\sigma_{irp}^2+\sigma_e^2$
10. C X I X R	$n\sigma_{cir}^2+\sigma_{irp}^2+\sigma_e^2$
11. I X R X P	$\sigma_{irp}^2+\sigma_e^2$
12. Error	σ_e^2

Setting the actual MS's of Table 12 equal to their respective expected MS's yields 12 linear equations whose solutions are given in Table 14. For illustrative purposes it was assumed that n = 2, c = 2, i = 4, and r = 6 which were the same as the values in Table 10. As a result,

$$\sigma_T^2 = (6\cdot4)^2\sigma_c^2 = 576\sigma_c^2,$$

Table 14

Estimation of Variance Components for a Three Factor Design with Repeated Measures on Two Factors

1. σ_c^2 (=) $\frac{1}{nir}(s_p^2+s_{ci}^2-s_{ip}^2+s_{cr}^2-s_{cir}^2+s^2)$ (=) $\frac{1}{48}(s_p^2+s_{ci}^2-s_{ip}^2+s_{cr}^2-s_{rp}^2-s_{cir}^2+s^2)$

2. σ_p^2 (=) $\frac{1}{ir}(s_p^2-s_{ip}^2-s_{rp}^2+s^2)$ (=) $\frac{1}{24}(s_p^2-s_{ip}^2-s_{rp}^2+s^2)$

3. σ_i^2 (=) $\frac{1}{ncr}(s_i^2-s_{ci}^2-s_{ir}^2+s_{cir}^2)$ (=) $\frac{1}{24}(s_i^2-s_{ci}^2-s_{ir}^2+s_{cir}^2)$

4. σ_{ci}^2 (=) $\frac{1}{nr}(s_{ci}^2-s_{ip}^2-s_{cir}^2+s^2)$ (=) $\frac{1}{12}(s_{ci}^2-s_{ip}^2-s_{cir}^2+s^2)$

5. σ_{ip}^2 (=) $\frac{1}{r}(s_{ip}^2-s^2)$ (=) $\frac{1}{6}(s_{ip}^2-s^2)$

6. σ_r^2 (=) $\frac{1}{nci}(s_r^2-s_{cr}^2-s_{ir}^2+s_{cir}^2)$ (=) $\frac{1}{16}(s_r^2-s_{cr}^2-s_{ir}^2+s_{cir}^2)$

7. σ_{cr}^2 (=) $\frac{1}{ni}(s_{cr}^2-s_{rp}^2-s_{cir}^2+s^2)$ (=) $\frac{1}{8}(s_{cr}^2-s_{rp}^2-s_{cir}^2+s^2)$

8. σ_{rp}^2 (=) $\frac{1}{i}(s_{rp}^2-s^2)$ (=) $\frac{1}{4}(s_{rp}^2-s^2)$

9. σ_{ir}^2 (=) $\frac{1}{nc}(s_{ir}^2-s_{cir}^2)$ (=) $\frac{1}{4}(s_{ir}^2-s_{cir}^2)$

10. σ_{cir}^2 (=) $\frac{1}{n}(s_{cir}^2-s^2)$ (=) $\frac{1}{2}(s_{cir}^2-s^2)$

11. σ^2 (=) s^2

$$\sigma_x^2 = (6\cdot4)(6\cdot4\sigma_c^2+6\cdot4\sigma_p^2+6\sigma_{ci}^2+6\sigma_{ip}^2+4\sigma_{rp}^2+\sigma_{cir}^2+\sigma^2)$$
$$+24(24\sigma_c^2+24\sigma_p^2+6\sigma_{ci}^2+6\sigma_{ip}^2+4\sigma_{cr}^2+4\sigma_{rp}^2+\sigma_{cir}^2+\sigma^2)$$

Therefore, $r=24\sigma_c^2/(24\sigma_c^2+24\sigma_p^2+6\sigma_{ci}^2+6\sigma_{ip}^2+4\sigma_{cr}^2+4\sigma_{rp}^2+\sigma_{cir}^2+\sigma^2)$

It will be noted that in all the designs discussed up to now the situation factor has been equal to one and therefore the qjt value which was to be multiplied by the σ_c^2 term to yield the numerator of the reliability estimate has essentially been equal to $qj\sigma^2$. The fact that $t = 1$ was also taken into account where necessary in the denominator as well.

It was pointed out in the discussion of the three factors with one repeated measure that the design was also applicable to cases in which the observer or teacher factor were repeated. The same reasoning obtains for the design with two repeated measures. That is, it is possible that teachers and items or teachers and observers rather than items and observers be considered the repeated factors. This is only possible if each teacher is observed more than once. If such is the case, the value of t in the calculation of r would obviously not be equal to one.

Inherent in the preceding idea of treating the teacher factor as a repeated measure is the assumption that the visit or situation factor need not be treated as a separate variable but may be subsumed under the teacher factor. However, it is possible to treat the teacher factor as a non-repeated measure by introducing a situation factor for the case in which the teacher was visited more than once. In such a situation, assuming the same observer and items, one may treat the experiment as a four way design with two repeated measures. The observer and item factors, the repeated measures, may be considered subsumed under the teacher and situation variables, respectively. The linear model may then be given as

$$\begin{aligned} X_{ijklm} = {} & u + A_i + B_j + AB_{ij} + P_{m(ij)} + C_k + AC_{ik} + BC_{jk} \\ & + ABC_{ijk} + CP_{km(ij)} + D_e + AD_{ie} + BD_{jl} + ABD_{ijl} \\ & + DF_{em(ij)} + CD_{kl} + ACD_{ikl} + BCD_{jkl} + ABCD_{ijkl} \\ & + CDP_{klm(ij)} + \text{error} \end{aligned}$$

where A, B, C, D are the teacher, recorder, situation and item factors, respectively.

The expected MS's for this model, assuming n subscores, are given in Table 15. Assuming that classes, recorders, situations, and items are random factors, D_c, D_r, D_s, and D_i all become equal to one. Therefore, lines 19 and 20 of Table 15 may be pooled to form a new "residual" which may serve as the error term throughout the design. The results of this change are given in Table 16. Table 17 gives the degrees of freedom for the design. The solution of the linear equations resulting from setting the observed MS's equal to their respective expected MS's are given in Table 18. Based on the estimated variance components,

$$\sigma_T^2 = (6 \cdot 4 \cdot 3)^2 \sigma_c^2$$

Table 15

Four Factor Design with Two Repeated Measures

	Source	E(MS)
1.	Class (C)	$nrsi\sigma_c^2+nsiD_r\sigma_{cr}^2+si\sigma_p^2+nriD_s\sigma_{cs}^2+niD_rD_s\sigma_{crs}^2+iD_s\sigma_{sp}^2+nrsD_i\sigma_{ci}^2+nsD_rD_i\sigma_{cri}^2+sD_i\sigma_{ip}^2+nrD_sD_i\sigma_{csi}^2+nD_rD_sD_i\sigma_{crsi}^2+D_sD_i\sigma_{sip}^2+\sigma_e^2$
2.	Recorder (R)	$ncsi\sigma_r^2+nsiD_c\sigma_{cr}^2+si\sigma_p^2+nciD_s\sigma_{rs}^2+niD_cD_s\sigma_{crs}^2+iD_s\sigma_{sp}^2+ncsD_i\sigma_{ri}^2+nsD_cD_i\sigma_{cri}^2+sD_i\sigma_{ip}^2+ncD_sD_i\sigma_{rsi}^2+nD_cD_sD_i\sigma_{crsi}^2+D_sD_i\sigma_{sip}^2+\sigma_e^2$
3.	C X R	$nsi\sigma_{cr}^2+si\sigma_p^2+niD_s\sigma_{crs}^2+iD_s\sigma_{sp}^2+nsD_i\sigma_{cri}^2+sD_i\sigma_{ip}^2+nD_rD_s\sigma_{crsi}^2+D_sD_i\sigma_{sip}^2+\sigma_e^2$
4.	S.W.g. (P)	$si\sigma_p^2+iD_s\sigma_{sp}^2+sD_i\sigma_{ip}^2+D_sD_i\sigma_{sip}^2+\sigma_e^2$
5.	Situation (S)	$ncri\sigma_s^2+nri\sigma_{cs}^2+nciD_r\sigma_{rs}^2+niD_cD_r\sigma_{crs}^2+s\sigma_{sp}^2+ncrD_i\sigma_{si}^2+nrD_cD_i\sigma_{csi}^2+ncD_rD_i\sigma_{rsi}^2+nD_cD_rD_i\sigma_{crsi}^2+D_i\sigma_{sip}^2+\sigma_e^2$
6.	C X S	$ri\sigma_{cs}^2+niD_r\sigma_{crs}^2+i\sigma_{sp}^2+nrD_i\sigma_{csi}^2+nD_rD_i\sigma_{crsi}^2+D_i\sigma_{sip}^2+\sigma_e^2$
7.	R X S	$nci\sigma_{rs}^2+niD_c\sigma_{crs}^2+i\sigma_{sp}^2+ncD_i\sigma_{rsi}^2+nD_cD_i\sigma_{crsi}^2+D_i\sigma_{sip}^2+\sigma_e^2$
8.	C X R X S	$niD_s\sigma_{crs}^2+iD_s\sigma_{sp}^2+nD_sD_i\sigma_{crsi}^2+D_sD_i\sigma_{sip}^2+\sigma_e^2$
9.	S X P	$i\sigma_{sp}^2+D_i\sigma_{sip}^2+\sigma_e^2$
10.	Item (I)	$ncrs\sigma_i^2+nrsD_c\sigma_{ci}^2+nsD_cD_r\sigma_{cri}^2+s\sigma_{ip}^2+ncrD_s\sigma_{si}^2+nrD_cD_s\sigma_{csi}^2+ncD_rD_s\sigma_{rsi}^2+nD_cD_rD_s\sigma_{crsi}^2+D_s\sigma_{sip}^2+\sigma_e^2$

Table 15 (continued)

Source	E(MS)
11. C X I	$nrs\sigma^2_{ci}+nsD_r\sigma^2_{cri}+s\sigma^2_{ip}+ncD_s\sigma^2_{rsi}+nD_cD_s\sigma^2_{crsi}+D_r\sigma^2_{sip}+\sigma^2_e$
12. R X I	$ncs\sigma^2_{ri}+nsD_c\sigma^2_{cri}+s\sigma^2_{ip}+ncD_s\sigma^2_{rsi}+nD_cD_s\sigma^2_{crsi}+D_r\sigma^2_{sip}+\sigma^2_e$
13. C X R X I	$ns\sigma^2_{cri}+s\sigma^2_{ip}+nD_s\sigma^2_{crsi}+D_s\sigma^2_{sip}+\sigma^2_e$
14. I X P	$s\sigma^2_{ip}+D_s\sigma^2_{sip}+\sigma^2_e$
15. S X I	$ncr\sigma^2_{si}+nrD_c\sigma^2_{csi}+ncD_r\sigma^2_{rsi}+nD_cD_r\sigma^2_{crsi}+\sigma^2_{sip}+\sigma^2_e$
16. C X S X I	$nr\sigma^2_{csi}+nD_r\sigma^2_{crsi}+\sigma^2_{sip}+\sigma^2_e$
17. R X S X I	$nc\sigma^2_{rsi}+nD_c\sigma^2_{crsi}+\sigma^2_{sip}+\sigma^2_e$
18. CXRXSXI	$n\sigma^2_{crsi}+\sigma^2_{sip}+\sigma^2_e$
19. S X I X P	$\sigma^2_{sip}+\sigma^2_e$
20. Error	σ^2_e

Table 16

Four Random Factors Design with Two Repeated Measures

	Source	E(MS)
1.	C	$nrsi\sigma^2_c+nsi\sigma^2_{cr}+si\sigma^2_p+nri\sigma^2_{cs}+ni\sigma^2_{crs}+i\sigma^2_{sp}+nrs\sigma^2_{ci}+ns\sigma^2_{cri}+s\sigma^2_{ip}+nr\sigma^2_{csi}+n\sigma^2_{crsi}+\sigma^2$
2.	R	$ncsi\sigma^2_r+nsi\sigma^2_{cr}+si\sigma^2_p+nci\sigma^2_{rs}+ni\sigma^2_{crs}+i\sigma^2_{sp}+ncs\sigma^2_{ri}+ns\sigma^2_{cri}+s\sigma^2_{ip}+nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$
3.	CXR	$nsi\sigma^2_{cr}+si\sigma^2_p+ni\sigma^2_{crs}+i\sigma^2_{sp}+ns\sigma^2_{cri}+s\sigma^2_{ip}+n\sigma^2_{crsi}+\sigma^2$
4.	P	$si\sigma^2_p+i\sigma^2_{sp}+s\sigma^2_{ip}+\sigma^2$
5.	S	$ncri\sigma^2_s+nri\sigma^2_{cs}+nci\sigma^2_{rs}+ni\sigma^2_{crs}+i\sigma^2_{sp}+ncr\sigma^2_{si}+nr\sigma^2_{csi}+nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$
6.	CXS	$nri\sigma^2_{cs}+ni\sigma^2_{crs}+i\sigma^2_{sp}+nr\sigma^2_{csi}+n\sigma^2_{crsi}+\sigma^2$
7.	RXS	$nci\sigma^2_{rs}+ni\sigma^2_{crs}+i\sigma^2_{sp}+n\sigma^2_{crsi}+\sigma^2$
8.	CXRXS	$ni\sigma^2_{crs}+i\sigma^2_{sp}+n\sigma^2_{crsi}+\sigma^2$
9.	SXP	$i\sigma^2_{sp}+\sigma^2$
10.	I	$ncrs\sigma^2_i+nrs\sigma^2_{ci}+ns\sigma^2_{cri}+s\sigma^2_{ip}+ncr\sigma^2_{si}+nr\sigma^2_{csi}+nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$
11.	CXI	$nrs\sigma^2_{ci}+ns\sigma^2_{cri}+s\sigma^2_{ip}+nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$
12.	RXI	$ncs\sigma^2_{ri}+ns\sigma^2_{cri}+s\sigma^2_{ip}+nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$
13.	CXRXI	$ns\sigma^2_{cri}+s\sigma^2_{ip}+n\sigma^2_{crsi}+\sigma^2$
14.	IXP	$s\sigma^2_{ip}+\sigma^2$
15.	SXI	$ncr\sigma^2_{si}+nr\sigma^2_{csi}+nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$

Table 16 (continued)

	Source	E(MS)
16.	CXSXI	$nr\sigma^2_{csi}+n\sigma^2_{crsi}+\sigma^2$
17.	RXSXI	$nc\sigma^2_{rsi}+n\sigma^2_{crsi}+\sigma^2$
18.	CXRXSXI	$n\sigma^2_{crsi}+\sigma^2$
19.	Residual	σ^2

Table 17

Degrees of Freedom for a Four Factor Design with Two Repeated Measures

Source	d.f.	Actual (MS)
Between	ncr - 1	
1. C	c - 1	s_c^2
2. R	r - 1	s_r^2
3. CXR	(c - 1)(r - 1)	s_{cr}^2
4. P	cr(n - 1)	s_p^2
Within	ncr(si - 1)	
5. S	s - 1	s_s^2
6. CXS	(c - 1)(s - 1)	s_{cs}^2
7. RXS	(r - 1)(s - 1)	s_{rs}^2
8. CXRXS	(c - 1)(r - 1)(s - 1)	s_{crs}^2
9. SXP	cr(n - 1)(s - 1)	s_{sp}^2
10. I	(i - 1)	s_i^2
11. CXI	(c - 1)(i - 1)	s_{ci}^2
12. RXI	(r - 1)(i - 1)	s_{ri}^2
13. CXRXI	(c - 1)(r - 1)(i - 1)	s_{cri}^2
14. IXP	cr(n - 1)(i - 1)	s_{ip}^2
15. SXI	(s - 1)(i - 1)	s_{si}^2
16. CXSXI	(c - 1)(s - 1)(i - 1)	s_{csi}^2
17. RXSXI	(r - 1)(s - 1)(i - 1)	s_{rsi}^2
18. CXRXSXI	(c - 1)(r - 1)(s - 1)(i - 1)	s_{crsi}^2
19. Residual	cr(n - 1)(s - 1)(i - 1)	s^2

$$= 5184\sigma_c^2$$

and
$$\sigma_X^2 = 6\cdot4\cdot3(6\cdot4\cdot3\ \sigma_c^2 + 4\cdot3\ \sigma_{cr}^2 + 4\cdot3\ \sigma_p^2 + 6\cdot4\ \sigma_{cs}^2 + 4\ \sigma_{cs}^2 + 4\ \sigma_{crs}^2 + 4\ \sigma_{sp}^2 + 6\cdot3\ \sigma_{ci}^2 + 3\ \sigma_{cri}^2 + 3\ \sigma_{ip}^2 + 6\ \sigma_{csi}^2 + \sigma_{crsi}^2 + \sigma^2)$$

$$\sigma_X^2 = 72(72\ \sigma_c^2 + 12\ \sigma_{cr}^2 + 12\ \sigma_p^2 + 24\ \sigma_{cs}^2 + 4\ \sigma_{crs}^2 + 4\ \sigma_{sp}^2 + 18\ \sigma_{ci}^2 + 30\ \sigma_{cri}^2 + 3\ \sigma_{ip}^2 + 6\ \sigma_{csi}^2 + \sigma_{crsi}^2 + \sigma^2).$$

Therefore,

$$r = 72\sigma_c^2 \ / \ D$$

where
$$D = 72\sigma_c^2 + 12\sigma_{cr}^2 + 12\sigma_p^2 + 24\sigma_{cs}^2 + 4\sigma_{crs}^2 + 4\sigma_{sp}^2 + 18\sigma_{ci}^2 + 3\sigma_{cri}^2 + 3\sigma_{ip}^2 + 6\sigma_{csi}^2 + \sigma_{crsi}^2 + \sigma^2$$

In all of the foregoing models the assumption was made that the factors were random variables. The reasons for this assumption were that (1) this is the more general case and may be applied to a specific instance of a fixed variable by dropping out the requisite number of terms through the use of the Medley and Mitzel (1963) "ground rules" and the fact that $D_x = 0$, and (2) that the assumption of a fixed variable tends to inflate the reliability calculation (Medley & Mitzel, 1963). Part of the data will be analyzed in Chapter III under the assumption that one of the variables was fixed and the resulting reliability estimate was considerably higher than it would have been had this assumption not been made.

Table 18

Variance Components for a Four Factor Design with Two Repeated Measures

1. σ_c^2 (=) $\frac{1}{nrsi}(s_c^2 - s_{cr}^2 - s_{ci}^2 - s_{cs}^2 + s_{crs}^2 + s_{cri}^2 + s_{rsi}^2 - s_{crsi}^2)$
2. σ_r^2 (=) $\frac{1}{ncsi}(s_r^2 - s_{ri}^2 - s_{rs}^2 - s_{cr}^2 + s_{crs}^2 + s_{cri}^2)$
3. σ_{cr}^2 (=) $\frac{1}{nsi}(s_{cr}^2 - s_p^2 + s_{ip}^2 + s_{sp}^2 - s_{crs}^2 - s_{cri}^2 + s_{crsi}^2 - s^2)$
4. σ_p^2 (=) $\frac{1}{si}(s_p^2 - s_{ip}^2 - s_{sp}^2 + s^2)$
5. σ_s^2 (=) $\frac{1}{ncri}(s_s^2 - s_{si}^2 - s_{rs}^2 - s_{cs}^2 + s_{crs}^2 + s_{csi}^2)$

Table 18 (continued)

6. σ^2_{cs} (=) $\frac{1}{nri}(s^2_{cs}-s^2_{crs}-s^2_{csi}+s^2_{crsi})$
7. σ^2_{rs} (=) $\frac{1}{nci}(s^2_{rs}-s^2_{crs})$
8. σ^2_{crs} (=) $\frac{1}{ni}(s^2_{crs}-s^2_{sp}-s^2_{crsi}+s^2)$
9. σ^2_{sp} (=) $\frac{1}{i}(s^2_{sp}-s^2)$
10. σ^2_{i} (=) $\frac{1}{ncrs}(s^2_{i}-s^2_{si}-s^2_{ci}+s^2_{rsi})$
11. σ^2_{ci} (=) $\frac{1}{nrs}(s^2_{ci}-s^2_{cri}-s^2_{rsi}+s^2_{crsi})$
12. σ^2_{ri} (=) $\frac{1}{ncs}(s^2_{ri}-s^2_{cri}-s^2_{rsi}+s^2_{crsi})$
13. σ^2_{cri} (=) $\frac{1}{ns}(s^2_{cri}-s^2_{ip}-s^2_{crsi}+s^2)$
14. σ^2_{ip} (=) $\frac{1}{s}(s^2_{ip}-s^2)$
15. σ^2_{si} (=) $\frac{1}{ncr}(s^2_{si}-s^2_{csi}-s^2_{rsi}+s^2_{crsi})$
16. σ^2_{csi} (=) $\frac{1}{nr}(s^2_{csi}-s^2_{crsi})$
17. σ^2_{rsi} (=) $\frac{1}{nc}(s^2_{rsi}-s^2_{crsi})$
18. σ^2_{crsi} (=) $\frac{1}{n}(s^2_{crsi}-s^2)$
19. σ^2 (=) s^2

CHAPTER III
Analysis of the Results of the Investigation

The SUTEC Observation data were subjected to two analyses. The first analysis dealt with the reliability of three of the items which comprised the schedule while the second analysis dealt with the reliability estimate for the entire schedule.

It was the purpose of the present section to develop applications of sOme of the ideas discussed in Chapter II. Thus, this chapter addresses itself mainly to the last four questions proposed at the outset of the investigation, the first three questions having been discussed in Chapter II. .

Reliability of Individual Items

As was pointed out in the procedure section of the preceding chapter, seven observers were present for each reliability visit to two teachers. However, only four of the same observers were present for both observations. The items observed were teacher mobility, involvement of children, and irrelevant acts. Because these observations were carried out while the schedule was being devised, it was decided that an ANOVA and a reliability coefficient would be calculated for each item rather than for the entire schedule.

The analyses were carried out by taking the general paradigm and applying it to this specific case. The model, in which all variables actually represent deviations from their respective means, was

$$X_{ijk} = C_i + O_j + I_{ij} + e_{ijk}$$

where C_i represented the deviation associated with teacher i, O_j the deviation associated with observer j, I_{ij} the interaction between teachers and observers, and e_{ijk} the "error" or residual term. Upon taking mathematical expectations, assuming infinite populations of teachers and observers, the result was

$$\sigma_x^2 = \sigma_c^2 + \sigma_o^2 + \sigma_{co}^2 + \sigma^2$$

where σ_x^2 was the total variance for all the x observations, and the terms on the right of this equation were the respective population variances for teachers, observers, interaction, and residual. Because the ratio, $F = MS_{co}/MS_{error}$, which tested the hypothesis H_1: $\sigma_{co}^2 = 0$ had a value less than one for mobility, involvement and irrelevant acts, the interaction and error terms were pooled to form a new residual term. The resulting ANOVA

Table 19
Analysis of Variance of the SUTEC Observation Team Data

			Observed MS					
Source	df	E(MS)	Mobility	F	Involv.	F	Irr. Acts	F
Class	1	$7\sigma_c^2+\sigma^2$	4.95	6.05*	38.79	16.13**	380.64	53.42**
Observation	6	$2\sigma_o^2+\sigma^2$	2.08	2.55	6.86	2.85	48.32	6.78*
Residual	6	σ^2	.82		2.41		7.13	

*$p < .05$
**$p < .01$

is given in Table 20. In both Tables 19 and 20, σ^2, σ_o^2, and σ_c^2 are the expected variances of the residual or error term, observers, and teachers or classes, respectively.

Table 20
Estimation of Variance Components of the SUTEC Observation Team Data

	E(MS)	Mobility	Involvement	Irrelevant Acts
σ_c^2 (=)[a]	$1/7\ (s_c^2 - s^2)$[b]	.59	5.20	53.36
σ_o^2 (=)	$1/2\ (s_o^2 - s^2)$	.63	2.23	20.77
σ^2 (=)	s^2	.82	2.41	7.13

[a] The symbol "(=)" is to read " is estimated by."
[b] The s^2 terms denote the actual mean squares.

The overall reliability coefficient was computed by using the formula

$$R = \sigma_T^2 \,/\, \sigma_X^2$$

where σ_T^2 and σ_X^2, the true and total variances, respectively, were defined by Medley and Mitzel (1963). Here,

$$\sigma_T^2 = (qjr)^2\ (\sigma_c^2) \text{ and } \sigma_X^2 = (qjr\ \sigma_c^2 + qr\ \sigma_o^2 + \sigma^2)$$

where q is the number of observation records, r the number of situations, and j the number of items. Therefore, here, $\sigma_T^2 = (7\cdot1\cdot1)^2\ \sigma_c^2 = 49\ \sigma_c^2$ and

$$\sigma^2 = (7\cdot1\cdot1)\ (7\cdot1\cdot1\ \sigma_c^2 + 7\cdot1\cdot\sigma_c^2 + \sigma_c^2)$$
$$= 49\ \sigma_c^2 + 49\ \sigma_o^2 + 7\sigma^2$$

If the data are analyzed only for the four observers who were present during both reliability visits the observer factor must be treated as a "fixed" rather than a "random" factor. Accordingly, the σ_o^2 component of σ_X^2 is zero, and $\sigma_{X2}^2 = (qjr)(qjr\ \sigma_c^2 + \sigma^2)$. Furthermore, the hypothesis H_1: $\sigma_o = 0$, tested by the ratio $\underline{F} = MS_o/MS_{error}$, yielded values of one or less for all three items and therefore this factor was pooled with the error term. The resulting ANOVA and estimation of variance components for these data are given in Tables 21 and 22, respectively.

Table 21
ANOVA for the Four Observers Present
During Both Observations

			Observed MS					
Source	df	E(MS)	Mobility	F	Involv.	F	Irr. Acts	F
Class	1	$4\sigma_c^2+\sigma^2$	6.13	6.39*	21.13	6.76*	180.50	10.08*
Residual	6	σ^2	.96		3.13		17.92	

*p <.05

Table 22
Estimation of Variance Components for the Four Observer Present During Both Visits

	E(MS)	Mobility	Involvement	Irrelevant Acts
σ_c^2	(=) $1/4\ (s_c^2 - s^2)$	1.19	4.50	40.65
σ^2	(=) s^2	.96	3.13	17.92

The appropriate values for σ_T^2 and σ_X^2 were calculated and are given in Table 23. These values were used to calculate the overall reliability coefficients for the entire observation team. These were r (mobility) = .72, r (involvement) = .67, r (irrelevant acts) = .69 and the corresponding coefficients for the four observers present during both observations were r (mobility) = .84, r (involvement) = .85, r (irrelevant acts) = .90. Clearly, then, one way to increase reliability would be to maintain the same observers throughout--a finding in complete agreement with common sense and the previously cited literature.

Table 23
Variances and Correlations for the Entire Observation Team and the Four Observers Present During Both Observations

	Mobility		Involvement		Irrelevant Acts	
Variance	Team	4 Observers	Team	4 Observers	Team	4 Observers
σ_T^2	26.41	20.67	254.68	72.00	2664.50	650.33
σ_X^2	36.57	24.50	380.58	84.50	3848.37	722.00
R	.72	.84	.67	.85	.69	.90

The calculated r's indicated that 28%, 33% and 31% of the variance was attributed to factors other than teachers for mobility, involvement, and irrelevant acts, respectively. By comparing the observer and residual variances it was found that 12.2% and 15.8, 15.9% and 17.1%, and 20.4% and 10.6% of the variances were due to observers and residual or errors for mobility, involvement, and irrelevant acts respectively.

The finding that the variances due to different teachers ranged from .67 to .72 indicated that the observation schedule differentiated between teachers on the variables investigated. Furthermore, in two of the three cases less than 16% of the variance was due to observers while in the third case approximately 20% of the variance was due to observers. This latter finding indicated a need for intensifying the training procedures of the observers on this factor, or the sharpening of the definition of this variable, or both.

As was indicated in Chapter II, fixed factors tend to inflate the reliability estimate and the average increase in r for the four observer case over the seven observer situation was .17. Besides the obvious rationalization, this was due to the fact that the denominator of r decreased more quickly as a result of the removal of the "fixed" factor. However, why irrelevant acts, the most subjective category yielded the highest r for the four observer situation still remains to be investigated.

Reliability of the Entire Schedule

Since each teacher was observed by an observer team peculiar to himself, the model was considered a partially hierachical design. That is, each observer team had the same number of observers but not necessarily the same observers and therefore the observer team factor was nested under the teacher factor. If teachers were factor A, observers factor B, and items factor C, B would be nested under A. Assuming that there were n scores on each item for each teacher per observer the sources of variation, degrees of freedom, and expected mean squares were as given in Table 24 (Winer, 1962) where p, q, and r were the numbers of teachers, observers, and items respectively.

The D_p, D_q, and D_r terms are equal to $1-p/P$, $1-q/Q$, $1-r/R$, respectively, where the p and P, q and Q, and r and R are the sample and population parameters of teachers, observers, and items, respectively. Each of these D's is either 0 or 1 depending on whether the corresponding factor is fixed or random.

As was pointed out by Medley and Mitzel (1963), the assignment of a variable as fixed tends to reduce the error of measurement and hence inflate the reliability. Therefore, the assumption that a variable is fixed should be based on sound reasons. A rule of thumb for selecting which factors are fixed and which are random is to decide whether other elements comprising the factor might have been used, and if so, then the factor is random (Medley & Mitzel, 1963). For example, if no observers other than the ones actually employed could have been used satisfactorally, then the observer factor would be fixed. Since there are always other teachers and observers available, theoretically anyway, these factors are usually considered random factors. These ideas are consonant with the definitions given in Chapter II of this investigation.

Table 24

Sources of Variation, Degrees of Freedom, and Expected Mean Squares for an ANOVA Design with Factor B Nested Under Factor A

Source of Variation	df	E(MS)
A	p-1	$nqr\sigma_a^2+nrD_q\sigma_b^2+nqD_r\sigma_{ac}^2$ $+nD_qD_r\sigma_{bc}^2+\sigma^2$
B W. A	p(q-1)	$nr\sigma_b^2+nD_r\sigma_{bc}^2+\sigma_e^2$
C	r-1	$npq\sigma_c^2+nqD_p\sigma_{ac}^2+nD_q\sigma_{bc}^2+\sigma_e^2$
AC	(p-1)(r-1)	$nq\sigma_{ac}^2+nD_q\sigma_{bc}^2+\sigma_e^2$
(B W.A)XC	p(q-1)(r-1)	$n\sigma_{bc}^2+\sigma_e^2$
Within	pqr(n-1)	σ_e^2

More precisely, as p, q, and r, the number of the sample elements, approach the values of P, Q, and R the number of elements in the population, the ratios p/P, q/Q, and r/R approach a value of one and therefore D_p, D_q, and D_r approach zero. If zeros are substituted for the D's the number of factors contained in the expected mean squares shrink and thus the reliability is increased because the denominator of the fraction which defines the reliability coefficient is decreased.

The model is also applicable even when there is only one score per item per observer for each teacher. In this case the model is the same as in Table 24 with n=1 and the within source of variation removed. If all factors are random and ones are substituted for the D's the model now yields an error term of $\sigma_{bc}^2+\sigma_e^2$ (Winer, 1962). The remaining expected mean square values follow in a similar fashion. To simplify the model still further the Medley and Mitzel (1963) procedure may be utilized. According to this procedure, the last term in the source of variation column, the residual, is considered to be the error term and is denoted by σ_e^2 rather than $\sigma_{bc}^2+\sigma_e^2$. The simplification of the error term and the substitution of ones for the n and the D's result in the expected mean squares shown in Table 25.

The only major difference between the Winer (1962) and Medley and Mitzel (1963) approach occurs in the F ratio testing the main effects of Factor A. This particular F ratio utilizes the nested factor B as its denominator, and has a larger expected mean square term in the simplified version than is called for by Winer (1962). The difference between the models is due to the σ^2_{bc} term. This therefore means that a significant F ratio testing the hypothesis $\sigma^2_a = 0$ in the simplified version would certainly be significant according to Winer (1962). Since the other two F ratios testing the hypotheses $\sigma^2_c = 0$ and $\sigma^2_{ac} = 0$ use the residual expected mean square as denominators, both the Medley and Mitzel (1963) and Winer (1962) approaches yield the same F values in these two cases.

Table 25
ANOVA Design with Factor B Nested Under Factor A, All Factors Random, and n = 1

Source of Variation	df	E(MS)
A	p-1	$qr\sigma^2_a+r\sigma^2_{b(a)}+q\sigma^2_{ac}+\sigma^2$
B W.A	p(q-1)	$r\sigma^2_{b(a)}+\sigma^2$
C	r-1	$pq\sigma^2_c+q\sigma^2_{ac}+\sigma^2$
AC	(p-1)(r-1)	$q\sigma^2_{ac}+\sigma^2$
Residual	p(q-1)(r-1)	σ^2

There are actually two homogeneity assumptions implied by the model. The first is that the source of variation due to B(A) represents the pooled variation of observers within teachers. The second results from the fact that the residual term is actually the B(A)XC interaction term and represents the pooling of different sources of variations. The homogeneity assumption here is equivalent to the assumption that the correlation between items is constant within each of the teachers.

Three teachers were observed once through a one way glass by three different observer teams. Each observer team contained seven members, but some of the observers were not the same throughout all the observations and therefore the teams were considered different.

In line with the earlier discussion of random and fixed variables, the teacher and observer factors were considered random factors, but because the observers were instructed to disregard all behavior other than those on the observation schedule the items were fixed. Accordingly, the σ^2_{ac} term in the first and third lines of Table 25 were dropped from the expected mean squares for teachers and items, respectively. The actual and expected mean squares for this specific situation in which p=3, q=7, and r=7 are given in Table 26.

Table 26
Analysis of Variance of an Observation Schedule
Containing Seven Items and Using
Three Observer Teams and Three Teachers

Source of Variation	df	E(MS)	Observed (MS)
A (Teachers)	2	$49\sigma^2_a+\sigma^2_{b(a)}+\sigma^2$	$s^2_a = 42.05$
B(A) (Observer within Teachers	18	$7\sigma^2_{b(a)}+\sigma^2$	$s^2_{b(a)} = 6.66$
C (Items)	6	$21\sigma^2_c+\sigma^2$	$s^2_c = 340.20$
AC	12	$7\sigma^2_{ac}+\sigma^2$	$s^2_{ac} = 56.42$
Residual	108	σ^2	$s^2 = 2.90$

The general set of linear equations which must be solved to find the estimated variance components is constructed by setting the estimated mean square terms equal to their corresponding observed mean squares. The resulting linear equations are then solved simultaneously. Table 27 gives the particular set of linear equations for the specific case listed in Table 26 and the resulting estimated values of the variances for each factor.

The three hypotheses $\sigma^2_a = 0$, $\sigma^2_c = 0$, $\sigma^2_{ac} = 0$ were all rejected because their respective F ratios,

$$F_a = MS_a / MS_{b(a)} = 6.31,$$

$$F_c = MS_c / MS_{residual} = 117.38,$$

$$F_{ac} = MS_{ac} / MS_{residual} = 19.47,$$

were all significant at the .01 level. The appropriate df's are given in Table 26. The rejection of these three hypotheses indicated that the scale does differentiate between the teachers and the items, and that there was a significant interaction between these two non nested factors.

Table 27
Estimation of Variance Components for an Observation Schedule Containing Seven Items and Using Three Observer Teams and Three Teachers

$$\sigma_a^2 \quad (=) \quad \frac{1}{49}(s_a^2 - s_{b(a)}^2) = .72$$

$$\sigma_{b(a)}^2 \quad (=) \quad \frac{1}{7}(s_{b(a)}^2 - s^2) = .54$$

$$\sigma_c^2 \quad (=) \quad \frac{1}{21}(s_c^2 - s^2) = 16.06$$

$$\sigma_{ac}^2 \quad (=) \quad \frac{1}{7}(s_{ac}^2 - s^2) = 7.65$$

$$\sigma^2 \quad (=) \quad s^2 = 2.90$$

The overall reliability coefficient (Medley & Mitzel, 1963) is equal to

$$R_{xx} = \sigma_T^2 / \sigma_X^2$$

Here, $\sigma_T^2 = (qr)^2 \sigma_a^2 = (7 \cdot 7)^2 \sigma_a^2 = 49^2(.7222) = 1734.00,$

and $\sigma_X^2 = qr\,(qr\sigma_a^2 + r\sigma_{b(a)}^2 + q\sigma_{ac}^2 + \sigma^2)$

$$= (7 \cdot 7)\,[(7 \cdot 7)\,(.7222) + 7(.5376) + 7(7.6460) + 2.8983] = 4682.99.$$

Therefore, $R_{xx} = 1734.00/4682.99 = .37$

The .37 reliability coefficient indicated that 37% of the variance was attributable to the teacher factor and 63% of the variance was due to the items, interactions, and residual factors. An examination of the ratio of the variances due to teachers and observers, the factor nested under teachers, indicated that 21.2% and 15.8% of the component of the total variance due to teachers was due to teachers and observers, respectively. A similar calculation for the other factors comprising the remaining 63% of the total variance yielded values of 38.0%, 18.1% and 6.9% for the items, interaction, and error or residual terms, respectively.

The proposed model did permit the partitioning of the variance associated with an observational schedule into its component parts and the calculation of an overall reliability coefficient. In the particular case to which the model was applied 75% of the variance was due to teachers and items, each of these two factors contributing equally to the total variance. Only 15.8% of the total variance was due to observers; the factor nested under teachers. These facts permit one to conclude that the variance due to different observers being used was considerably smaller than that due to the different teachers as they were observed on the various types of behavior represented by the items of the observational schedule.

That the items accounted for the single largest source of variance was probably due to the very different elements of behavior being observed. For example, materials present required very little judgment on the part of the observer, while involvement of children required a great deal of judgment.

As a result of this reliability study some confidence can be placed in the observation schedule's ability, as used by this observation team, to differentiate between teachers. A well trained team might therefore be used to observe teachers who were trained at various institutes or under different conditions at the same institution for the purposes of comparison. The data could then be analyzed and if differences existed, the superiority of one method of teacher training over another could be inferred.

The data could also have been analyzed using a repeated measures design as was pointed out earlier. This analysis yielded exactly the same information, resulting from the nested design used here. Verification is left as an exercise for the interested reader.

CHAPTER IV
Conclusions and Recommendations

Conclusions

This investigation sought to develop and apply analysis of variance techniques to the estimation of the reliability of observation schedules.

The investigation placed special emphasis on the different possible designs and the various administrative situations under which they might be applied. The application of the general model to a specific instance was then carried out.

The study was conducted with 10 recorders who observed five teachers through a one-way mirror and rated them on an observational schedule. This procedure was followed for each of three of the item categories comprising the schedule and for the entire schedule. In the first instance, two teachers were observed by teams of seven recorders. In the second situation, three teachers were observed by teams of seven recorders.

The materials used in this investigation was the SUTEC Observational Schedule which contained seven items. The observations for the estimation of reliability were all carried out at SUTEC.

Analysis of the data revealed that the overall reliability coefficient was .37 and that .72, .67, and .69 were the reliability coefficients for the mobility, involvement, and irrelevant acts items, respectively. When the observer factor was treated as a fixed factor the item reliabilities became .84, .85, and .90, respectively. Seventy-five per cent of the variance was accounted for by teachers and items for the overall reliability calculation, while approximately 70% of the variance was attributed to the teacher factor for the individual item reliabilities.

At this juncture, it must be pointed out that the application of the ANOVA technique to the SUTEC data does not even exhaust the few designs described earlier. Rather, this application was meant as an illustrative example of the wide range of possibilities which more accurate reliability calculations of observational data make possible.

Once reliable observations are possible, these types of data which may have been considered rather subjective will no longer be avoided by research workers. The respectability of observational data which may result has ramifications for a number of areas such as teacher supervision and training. However, any situation in which observations may be used is actually an area in which the reliability of the data may be calculated as indicated earlier in this paper. It may therefore be possible to utilize observational data in such disparate fields as educational psychology, industrial psychology, and social psychology.

The obvious areas of educational psychology such as teacher supervision and training have been stressed throughout this paper. Other aspects of school situations, particularly observations of children's behaviors during the teaching-learning situation as well as classroom and playground social interactions may now be studied.

Areas of industrial psychology, such as the behavior of workers under different conditions, market research, and behavior during labor disputes and labor negotiations may come under more rigorous study through the use of reliable observations. Some social psychologists might find observations to be a fruitful way of studying such diverse phenomena as mob reactions, school disorders, and the behavior of juries.

Clearly then, the work presented in this paper has ramifications for many fields. The application was specific to a pedagogical situation because the problem first came to the attention of the investigator in an educational context in which the data generated was related to an aspect of teacher training.

The major conclusions, presented within the limited scope of this investigation, were that different variance components models could be applied in different situations to estimate the reliability of either the entire observation schedule or parts of it, and that the items comprising the SUTEC schedule did differentiate fairly well between teachers.

Recommendations

The results of the present research prompted the following recommendations:

1. Construction of Computer programs to analyze observational data gathered under the various models described herein.

2. Extension of the models to situations in which an unequal number of observers, items, or situations were used without

requiring that some of the data be randomly discarded. This might be possible through an unweighted-means or a least-squares solutions analysis.

3. Investigation of the paradigms presented as a possible means of determining the homogeneity of the items comprising a schedule or proposed schedule.

4. Field testing of the different models simultaneously to permit comparison of the results. If the differences in the estimated reliabilities are slight, the simplest administrative procedures could then be adopted as the standard.

References

Bloom, R., & Wilensky, J. Four observation categories for rating teacher behavior. Journal of Educational Research, 1967, 60, 464-465.

Bobbitt, R. A., Gordon, B. N., & Jensen, G. D. Development and application of an observational method: Continuing reliability testing. Journal of Psychology, 1966, 63, 83-88.

Brown, B. B., Mendenhall, W., & Beaver, R. The reliability of observations of teacher's classroom behavior. Journal of Experimental Education, 1968, 36, 1-10.

Chapline, E. School University Teacher Education Center Technical Progress Report. September, 1968, Cooperative Research Project no. 5-0945, Contract no. OEC 1-6-050 945-1673, United States Office of Education.

Clifford, T. Simplified computational programming for analysis of variance designs with correlated observations. Psychological Bulletin, 1968, 69, 439-440.

Courson, C. C. The use of inference as a research tool. Educational & Psychological Measurement, 1965, 25, 1029-1038.

Denny, D. A. Identification of teacher-classroom variables facilitating pupil creative growth. American Educational Research Journal, 1968, 5, 365-383.

Flanders, N. A. Interaction analysis in the classroom: a manual for observers. Ann Arbor, Mich.: University of Michigan, 1960.

Furst, N., & Amidon, E. J. Teacher pupil interaction patterns in the elementary school. In E. J. Amidon & J. B. Hough (Eds.), Interaction analysis: theory, research, and application. Reading, Mass.: Addison-Wesley, 1967.

Katz, L. G., Peters, D. L., & Stein, N. S. Observing behavior in kindergarten and preschool classes. Childhood Education, 1968, 44, 400-405.

Maas, J. B. Patterned scaled expectation interview: reliability studies of a new technique. Journal of Applied Psychology, 1965, 49, 431-433.

McNemar, Q. Psychological statistics. New York: Wiley, 1962.

Medley, D. M. The use of orthogonal contrasts in the interpretation of records of verbal behaviors of classroom teachers. Research Memorandum RM-67-25. Princeton, N. J.: Educational Testing Service, 1967.

Medley, D. M., & Mitzel, H. E. Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. Journal of Experimental Education, 1958, 27, 23-35. (a)

Medley, D. M., & Mitzel, H. E. A technique for measuring classroom behavior. Journal of Educational Psychology, 1958, 49, 86-92. (b)

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 247-328.

Millman, J., & Glass, G. V. Rules of thumb for writing the Anova table. Journal of Educational Measurement, 1967, 4, 41-51.

Ojemann, R. H., & Snider, J. The effect of teaching program in behavioral science on changes in causal behavior scores. Journal of Educational Research, 1964, 57, 255-260.

Peng, K. C. The design and analysis of scientific experiments. Reading, Mass.: Addison-Wesley, 1967.

Rusch, R. R., Denny, D. A., & Ives, S. The development of a test of creativity in the dramatic arts: a pilot study. Journal of Educational Research, 1964, 57, 250-254.

Scott, W. A. Reliability of content analysis: the case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Seibel, D. W. Predicting the classroom behavior of teachers. Journal of Experimental Education, 1967, 36, 26-32.

Soar, R. S. An integrative approach to classroom learning. Philadelphia: Temple University, 1966.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Zunich, M. Child behaviors and parental attitudes. Journal of Psychology. 1966, 62, 41-46.

APPENDIX

School University Teacher Education Center

P.S. 76 and Queens College Education Department
36-36 Tenth Street
Long Island City, N. Y. 11106

School______________________ Teacher__________________

Grade_______________________ Date_____________________

Observer____________________ Time_____________________

Developed for use in the SUTEC project, November, 1967 by Elaine Chapline, Ph.D. and Theodore Abramson, M.S.

Attachment 1

SUTEC Observation

Number of children in class____

ROOM SKETCH

Front

Rear

Indicate positions of (W) windows, (D) door(s), (TD) teacher's desk, (CD) Children's desk, in groups, (SI) special interest areas.

Teacher mobility is indicated by marking teacher position on the room sketch during the second five minutes of each activity. Use an ordered pair to work each position.

i.e. 1,1=first activity, position one
1,2=first activity, position two, etc.
2,1=second activity, position one
2,2=second activity, position two, etc.
3,1=third activity, position one, etc.

Attachment 2

Involvement of Children

Activity

1. ______________________________
1 2 3

2. ______________________________
1 2 3

3. ______________________________
1 2 3

4. ______________________________
1 2 3

5. ______________________________
1 2 3

6. ______________________________
1 2 3

Scale

1. Uninvolved
2. Moderately involved
3. Highly involved

Attachment 3

Materials		Present	In Use
1. Chalkboard	1.		
2. Bulletin board(s)	2.		
3. Maps, charts, or pictures	3.		
4. Visual Aids (films, etc.)	4.		
5. Audio Aids (records, etc.)	5.		
6. Text	6.		
7. Library materials, magazines	7.		
8. Arts and crafts	8.		
9. Play materials (dolls, blocks, etc.)	9.		
10. Science equipment (fish tank, etc.)	10.		
11. Commercial supplemental materials (games, rex. sheets, workbooks, programmed materials, etc.)	11.		
12. Teacher made supplemental materials	12.		

Attachment 4

This observation of behavior should be used when the teacher is directing an activity for either the total class or a subgroup. Keep these tallies for 5 minutes, i.e. the third 5 minutes of an activity. Note if the time sample is other than 5 minutes.

Categories	Tallies					
	Activities					
	1	2	3	4	5	6
I						
II						
III						

Definitions of Categories:

I: A child talks or moves relevant to the activity without the teacher's direction or permission.

II: Teacher calls on child as a result of "hand raising" by child.

III: Teacher calls on child without prior "hand raising" by child.

Attachment 5

Child	No. of Irrelevant Acts	Sex	Totals
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			

9:45 AM
1:45 PM

Find the child nearest to you and observe him for 2 minutes. Record each irrelevant act with a tally. Find the third child from the one just observed and record his irrelevant actions for a two minute period. Continue, until six children have been observed, for a total of 12 minutes.

10:15 AM
2:15 PM

Continue from last child until six more children have been observed. Children may be repeated if it is their turn on the second go-around.